

Japanese Patent Laid-open No. 2003-131917



[Title of the Invention]

Storage System

[Abstract]

[Object]

An object of the present invention is to, in remote copies made among storage controllers that are located in N sites (the number of sites is in general three or more), manage differences that are used to quickly make remote-copy data coincide with one another among the remaining sites after an arbitrary site suffers from disaster.

[Solving Means]

An update history of a remote-copy range in a site, which is directly updated from a host computer, and an update history of a remote-copy range in another site, which is a copy destination of the remote-copy range in question, are shared between the sites in question. Writing of the update histories is triggered by a certain update from the host computer. When an arbitrary site suffers from disaster, contents of the remote-copy range are made coincide by copying, according to the update history, part of contents of the remote-copy range existing between arbitrary sites that do not suffer from disaster.

**BEST AVAILABLE COPY**

[Claims]

[Claim 1]

A storage system characterized in that:

in a remote-copy configuration so formed that storage systems are connected to one another through a network among sites, said sites each including a host computer and a storage system that is connected to the host computer, and data updated by the host computer is reflected for a storage system in another site,

an update history of a remote-copy specified range of a storage system, which is directly updated by the host computer, and an update history of a copy-to remote-copy specified range of a storage system in another site, which is a remote-copy destination of the remote-copy specified range, are shared between storage systems each including a remote-copy specified range and a copy-to remote-copy specified range, writing of the update histories being triggered by a certain update from the host computer, and

when an arbitrary site suffers from disaster, by copying part of contents of a remote-copy specified range or part of contents of a copy-to remote-copy specified range, which exist between arbitrary sites that do not suffer from disaster, according to the update history, the contents of the remote-copy specified range or the contents of the copy-to remote-copy specified range are made to coincide with

each other.

[Claim 2]

A storage system according to Claim 1, wherein:

areas to which the update history of the remote-copy specified range is written are provided in a multiplexed manner, and when the amount of updates from the host computer to the remote-copy specified range of the storage system, which is directly updated by the host computer, exceeds a threshold value, an update from the host computer, which triggers switching to a new update history, is determined, said determination being shared between storage systems each including a remote-copy specified range and a copy-to remote-copy specified range;

in each storage system, from a point of time at which updates to the remote-copy specified range or to the copy-to remote-copy specified range reach an update from the host computer, which triggers switching to a new update history, writing to the new update history is started; and

after checking that all storage systems, each of which includes a remote-copy specified range and a copy-to remote-copy specified range, have started writing of the new update history, writing of an existing update history is interrupted, and then the existing update history is initialized.

[Claim 3]

A storage system according to Claim 2, wherein:

said determination of the update from the host computer, which triggers switching to the new update history, is made also when a length of time elapsed after the last switching of the update history exceeds a threshold value.

[Claim 4]

A storage system according to Claim 2, wherein:

exchanges of information between the storage systems are made only with a direct copy source, and a direct copy destination, corresponding to the remote-copy specified range, or the copy-to remote-copy specified range, of the storage system in question.

[Claim 5]

A storage system according to Claim 2, wherein:

exchanges of information between the storage systems are made with said information being attached to remote-copy data to be transferred or received.

[Claim 6]

A storage system according to Claim 2, wherein:

when an arbitrary site suffers from disaster, an update history to be used to make contents of a remote-copy specified range or contents of a copy-to remote-copy specified range coincide between arbitrary sites that do not suffer from disaster is selected in both of the sites on the conditions that the update history is valid, and that

updates from the host computer which has started writing are the same; and

if there are a plurality of update histories each satisfying the conditions in both of the sites, an update history whose update from the host computer which has started the writing is the latest is selected.

[Claim 7]

A storage system according to Claim 2, wherein:

if it is detected that an arbitrary site has suffered from disaster, writing of an existing update history is continued, and at a point of time when the site which has suffered from the disaster is completely recovered to a normal state, the writing of the existing update history is interrupted, and then the existing update history is initialized.

[Claim 8]

A storage system according to Claim 2, wherein:

if it is detected that an arbitrary site has suffered from disaster, by use of areas for update histories from which an existing update history is excluded, the steps described in Claim 2 are executed between sites that do not suffer from disaster.

[Detailed Description of the Invention]

[0001]

[Technical Field to which the Invention Pertains]

The present invention relates to remote-copy functions used to duplex data between storage systems with a host computer not being involved. It relates more particularly to, in remote copies made among respective storage systems located in three or more sites, when an arbitrary site suffers from disaster, if the remote copying is continued among sites that do not suffer from the disaster, how to manage contents of a remote-copy specified range so that the contents are made to coincide with each other among the storage systems.

[0002]

[Prior Art]

In order to prevent data in a storage system from being lost due to disaster or the like, the data is duplexed in another storage system located in a remote place. Examples of this duplexing function include remote copy technologies.

[0003]

In the remote copy technologies, a target volume to be remote-copied is set in a primary storage controller connected to a host computer, and also in a secondary storage controller connected to the primary storage controller; and copies are continuously being executed so that contents of a primary volume of a primary storage device connected to the primary storage controller is always

made coincide with contents of a secondary volume of a secondary storage device connected to the secondary storage controller.

[0004]

The technology for writing data in duplication between different storage controllers with a host computer not being involved is disclosed in Japanese Patent Laid-open No. 11-85408. According to the above invention, a primary storage controller, which has received from a host computer write data to which write time is added, notifies the host computer that writing of the write data has been completed, and then transfers the write time and the write data to a secondary storage controller in the order of the write time. In a secondary storage controller, the write time and the write data which have been received from the primary storage controller are stored in a nonvolatile cache memory to ensure that data stored before this specified write time is reliable.

[0005]

The remote-copy operation is roughly classified into two: a synchronous remote copy for transferring write data to the secondary storage controller before notifying the host computer of the completion of a write request; an asynchronous remote copy for transferring write data to the secondary storage controller in asynchronization with the

write request after notifying the host computer of the completion of the write request. If a remote-copy function is used for applications such as a database, it is necessary to update a secondary volume in the same order as the update order in which a primary volume has been updated so as to ensure the update order of a log. In order to achieve the above, in the case of the asynchronous remote copy, every time a write is made from a host computer, a sequence number is added to write data in a primary storage controller. Then, the write is reflected in a secondary volume according to the sequence number in a secondary storage controller.

[0006]

In the conventional remote-copy operation described above, the remote-copy operation may be temporarily interrupted (for example, in a case where the remote-copy operation is interrupted by a link failure between the primary storage controller and the secondary storage controller, or the like). In such temporal interruption, with the objective of making contents of the secondary volume coincide with those of the primary volume as quickly as possible when the remote-copy operation is restarted, an update history may be written in the primary storage controller as difference information between the primary and secondary volumes. In the update history, updates from the host computer to the primary volume performed during the



interruption of the remote-copy operation are recorded. After that, when the remote-copy operation is restarted, by using the written update history to reflect the contents of the primary volume in the secondary volume, it is possible to make the contents of the secondary volume coincide with those of the primary volume.

[0007]

The difference information in the conventional remote copy usually means update locations from the host computer. More specifically, a location at which the primary volume is updated but the secondary volume is not updated is written. Therefore, the difference information is in general stored on the primary storage controller side.

[0008]

In recent years, with the increase in speed of networks, and with the decrease in cost of networks, costs of data transmission between remote locations are decreasing. For this reason, to increase the amount of transferred data in the remote-copy operation, and thereby to further increase the availability of remote-copy data, there is a demand that remote copies are made among N sites (the number of sites is three or more) so that data is kept duplexed even if one site suffers from disaster.

[0009]

As operational management of remote copies among N

sites, if one site among  $N$  sites suffers from disaster, it is desirable that a remote-copy configuration be made again by use of the remaining  $(N - 1)$  sites so as to continue the remote-copy operation. In this case, in order to continue the remote-copy operation, it is necessary to quickly make remote-copy data coincide between the remaining  $(N - 1)$  sites. In other words, it is desirable to perform difference management among  $(N - 1)$  sites so that the remote-copy data is quickly made coincide with one another by copying the differences.

[0010]

It is because if the difference management is not performed, there may arise a case where it becomes necessary to copy all remote-copy data from a copy source to a copy destination. In comparison with the copy of the differences, the copy of all remote-copy data takes a long time, and the availability of data is decreased while the copy is being executed. Therefore, it is not possible to make full use of the redundancy achieved by the number of sites.

[0011]

Here, as an example, on the assumptions that a remote-copy configuration is made with three sites, and that the conventional difference management method is applied to remote-copy operation, problems which arise are considered. Here, three sites are designated as A, B, C.

[0012]

First of all, if a flow of remote-copy data to be transferred among the sites are  $A \rightarrow B \rightarrow C$ , in the event that the site B suffers from disaster, it is desirable that differences be copied between a volume in the site A and a volume in the site C before restarting the remote-copy operation. Because the difference information in the conventional remote copy is in general stored on the primary storage controller side, difference management information between the site A and the site B is stored in the site A, and difference management information between the site B and the site C is stored in the site B. For this reason, if the site B suffers from disaster, the difference management information between the site B and the site C is lost. Accordingly, differences between the site A and the site C cannot be known.

[0013]

In addition, if a flow of remote-copy data to be transferred among the sites is  $A \rightarrow B$  and  $A \rightarrow C$ , in the event that the site A suffers from disaster, it is desirable that differences be copied between a volume in the site B and a volume in the site C before restarting the remote-copy operation. In this case, because difference management information between the site A and the site B and that between the site A and the site C are stored in the site A

that has suffered from the disaster, differences between the site B and the site C cannot be known.

[0014]

In order to cope with this problem, a method is conceivable in which difference management is performed beforehand between sites between which remote-copy data is not exchanged during the normal operation. In the above example, a difference between the site A and the site C is managed before the site B suffers from disaster; and a difference between the site B and the site C is managed before the site A suffers from disaster. However, if this method is used, the amount of difference information between sites which must be managed beforehand increases in proportion to the number of sites. Furthermore, it is not possible to manage differences in the event of unexpected circumstances including, for example, a case where two sites concurrently suffer from disaster.

[0015]

[Problems to be Solved by the Invention]

The conventional difference management method used for remote copies between two sites does not take difference management among N sites into consideration.

[0016]

Therefore, when remote copies are being made among N sites, if one site suffers from disaster, and consequently,

the remote-copy operation is continued among the remaining (N - 1) sites that do not suffer from disaster, there arises a problem in that a remote-copy configuration can be made again only after transferring all remote-copy data from a new copy-from site to a new copy-to site.

[0017]

In addition, if the conventional difference management method used for remote copies between two sites is applied beforehand to a pair of new sites when a remote-copy configuration is made again after a site has suffered from disaster, there arises a problem in that the amount of difference management information of each site increases with the increase in the number of sites. Further, if a plurality of sites suffer from disaster, there arises another problem in that differences cannot be managed.

[0018]

An object of the present invention is to, in remote copies made among storage systems that are located in N sites (the number of sites is in general three or more), after an arbitrary site suffers from disaster, manage differences that are used to quickly make remote-copy data coincide with each other among the remaining sites. Another object of the present invention is to keep the amount of information used to manage the differences constant without depending on the number of sites N among which remote copies

are made.

[0019]

[Means for Solving the Problems]

In order to achieve the above-mentioned objects, according to one aspect of the present invention, there is provided a storage system, wherein:

in a remote-copy configuration so formed that storage systems are connected to one another through a network among sites, said sites each including a host computer and a storage system that is connected to the host computer, and data updated by the host computer is reflected for a storage system in another site,

an update history of a remote-copy specified range of a storage system, which is directly updated by the host computer, and an update history of a copy-to remote-copy specified range of a storage system in another site, which is a remote-copy destination of the remote-copy specified range, are shared between storage systems each including a remote-copy specified range and a copy-to remote-copy specified range, writing of the update histories being triggered by a certain update from the host computer, and

when an arbitrary site suffers from disaster, by copying part of contents of a remote-copy specified range or part of contents of a copy-to remote-copy specified range, which exist between arbitrary sites that do not suffer from

disaster, according to the update history, the contents of the remote-copy specified range or the contents of the copy-to remote-copy specified range are made to coincide with each other.

[0020]

In addition, areas to which the update history of the remote-copy specified range is written are provided in a multiplexed manner, and when the amount of updates from the host computer to the remote-copy specified range of the storage system, which is directly updated by the host computer, exceeds a threshold value, an update from the host computer, which triggers switching to a new update history, is determined, the determination being shared between storage systems each including a remote-copy specified range and a copy-to remote-copy specified range; in each storage system, from a point of time at which updates to the remote-copy specified range or to the copy-to remote-copy specified range reach an update from the host computer, which triggers switching to a new update history, writing to the new update history is started; and after checking that all storage systems, each of which includes a remote-copy specified range and a copy-to remote-copy specified range, have started writing of the new update history, writing of an existing update history is interrupted, and then the existing update history is initialized.

[0021]

In addition, the determination of the update from the host computer, which triggers switching to the new update history, may be made also when a length of time elapsed after the last switching of the update history exceeds a threshold value.

[0022]

In addition, exchanges of information between the storage systems may also be made only with a direct copy source, and a direct copy destination, corresponding to the remote-copy specified range, or the copy-to remote-copy specified range, of the storage system in question.

[0023]

Further, the exchanges of information between the storage systems may also be made with said information being attached to remote-copy data to be transferred or received.

[0024]

Moreover, when an arbitrary site suffers from disaster, an update history to be used to make contents of a remote-copy specified range or contents of a copy-to remote-copy specified range coincide between sites that do not suffer from disaster is selected in both of the sites on the conditions that the update history is valid, and that updates from the host computer which has started writing are the same; and additionally, if there are a plurality of



update histories each satisfying the conditions in both of the sites, an update history whose update from the host computer which has started the writing is the latest may also be selected.

[0025]

Furthermore, an instruction which is used in the event of a failure in a remote-copy destination is given from a maintenance terminal, the instruction instructing to allow or disallow the execution of a remote copy to the other remote-copy destinations for each remote-copy destination of a remote-copy specified range or of a copy-to remote-copy specified range; and when a failure occurs in a remote-copy destination, remote copies from the storage system in question to the remote-copy destinations may be controlled according to the instruction.

[0026]

[Modes for Carrying out the Invention]

A first embodiment of the present invention will be described with reference to drawings below.

[0027]

Fig. 1 is a diagram illustrating how remote copies among N sites are made. In each site, each of a plurality of host computers 1, each of a plurality of storage systems 2, and the like, are connected to one another through each local network T2 such as a LAN (Local Area Network) or a SAN

(Storage Area Network). On the other hand, among the storage systems 2 in those sites, storage systems 2 between which a remote copy is made are connected to each other through a global network T1. In general, the global network T1 is a public telephone service, which is often leased from a communication service provider with charge. However, the present invention is not limited by configurations of the local networks T2, and a configuration of the global network T1.

[0028]

Target volumes to be remote-copied are roughly classified into: a top copy-from volume V1 that does not have a copy-from volume, and that has a copy-to volume; a middle copy-to volume V2 that has a copy-from volume and a copy-to volume; and an end copy-to volume V3 that has a copy-from volume, and that does not have a copy-to volume. The relationship among the target volumes to be remote-copied forms a tree shape with the top copy-from volume V1 being a vertex. Data to be remote-copied is transferred in succession from the top copy-from volume V1 to the end copy-to volume V3 through the middle copy-to volume V2.

[0029]

Fig. 1 illustrates a configuration in which a volume in the site 1 is specified as the top copy-from volume V1. In this configuration, a remote copy is made through the

middle copy-to volume V2 in the site 2 to the end copy-to volume V3 in the site (K - 1) and the end copy-to volume V3 in the site N, and also a remote copy is made through the middle copy-to volume V2 in the site K to the end copy-to volume V3 in the site (N -1).

[0030]

Fig. 2 is a diagram illustrating a configuration of the storage system 2. A storage controller 21 comprises components of: channel interfaces 3 each of which is connected to the host computer 1; disk interfaces 4 each of which is connected to a storage device 22; duplexed nonvolatile management information memories 6 for storing management information; and duplexed nonvolatile cache memories 5 for storing data. The above components are connected to one another through paths. Each of the channel interfaces 3 and each of the disk interfaces 4 are connected to the duplexed management information memories 6 and also to the duplexed cache memories 5. In addition, the storage controller 21 further comprises a maintenance terminal 23 used to instruct the storage controller 21 in question, and to display an internal state of the storage controller 21 in question.

[0031]

The channel interface 3 comprises: write processing 31 for handling a write request received from the host

computer 1, and a write request for a remote copy received from another storage system 2 that is a source of the remote copy; bit-map update processing 32 for updating a bit map corresponding to a range within which a write has been made; asynchronous transmission processing 33 for, if the own storage system 2 is a source of an asynchronous remote copy, transferring write data to said another storage system 2 as a copy destination in asynchronization with a write request; bit-map switching start processing 34 for triggering, from the storage system 2 having the top copy-from volume V1, switching of a bit map 809 covering all volumes that are copy destinations of the volume in question; bit-map switching processing 35 for receiving a bit-map switching request from the bit-map switching start processing 34, and then for switching the bit map; and difference copy processing 36 for, when reconfiguring a remote-copy configuration among remaining sites after suffering from disaster in an arbitrary site, making volume contents in each of the remaining sites coincident with those in the other remaining sites.

[0032]

The disk interface 4 comprises: asynchronous normalization processing 41 for, if the own storage system 2 is a destination of an asynchronous remote copy, normalizing remote-copy data that is temporarily saved as temporary data

at the time of writing from a copy source; and bit-map update processing 32 that is the same as the bit-map update processing 32 on the side of the channel interface 3.

[0033]

There are a plurality of pieces of processing for each kind of the above processing. As a unit of existence, the bit-map switching start processing 34, the bit-map switching processing 35, and the difference copy processing 36 may exist on a volume basis; or the asynchronous transmission processing 33 and the asynchronous normalization processing 41 may also exist for each write-sequence management information 62 described below.

[0034]

The cache memories 5 store data that is stored in the storage device 22, and that is read/written from/to the host computer 1. Each of the cache memories 5 can be broadly classified into a normal area and a temporary area. The reason why each cache memory 5 has the temporary area is that if the own storage system 2 is a source of an asynchronous remote copy, when untransferred remote-copy data is written to a copy destination again, it is necessary to save untransferred data before the write, or that if the own storage system 2 is a destination of an asynchronous remote copy, it is necessary to save remote-copy data existing after the remote-copy data is received from a copy

source before the remote-copy data is normalized by the asynchronous normalization processing 41.

[0035]

The management information memories 6 store management information required for the operation of the storage system 2. The management information includes remote-copy management information 61. The remote-copy management information 61 includes write-sequence management information 62 and volume management information 63.

[0036]

The write-sequence management information 62 is used to normalize remote-copy data in the write order received from the host computer 1 in an asynchronous remote copy. There are a plurality of pieces of the write-sequence management information 62. Each piece of write-sequence management information 62 includes: a write-sequence number counter 71 for writing the write order received from the host computer 1; a write-sequence management information entry 721 of a write-sequence management information entry area 72, which is assigned on a write data basis, and in which a write sequence number is saved so that the write sequence number is associated with write data; a transmission target write-sequence management information 73 for, in a copy source of an asynchronous remote copy, registering untransferred remote-copy data in a copy

destination; and normalization target write-sequence management information 74 for, in a copy destination of an asynchronous remote copy, registering remote-copy data that has been received from a copy source but has not yet been normalized.

[0037]

Fig. 3 is a diagram illustrating a structure of the write-sequence management information 62 and the relationship between the write-sequence management information 62 and the cache memory 5. The write-sequence management information entry 721 is assigned to each piece of write data. A location of the write data in the cache memory 5 is stored in the write-sequence management information entry 721. In addition, the transmission target write-sequence management information 73 and the normalization target write-sequence management information 74 may also be configured to have a queue structure for connecting the write-sequence management information entry 721.

[0038]

In addition, if it is necessary to ensure the write order received from the host computer 1 among a plurality of volumes, a write sequence number is assigned to write data in each of the plurality of volumes by use of the write-sequence number counter 71 of the same write-sequence

management information 62, and then the write sequence numbers are registered in the same transmission target write-sequence management information 73. In a copy destination, the write sequence numbers are registered in the same normalization target write-sequence management information 74, and normalizations are made according to the write sequence numbers.

[0039]

Returning to Fig. 2, the volume management information 63 is management information about a target volume to be remote-copied. There are a plurality of pieces of the volume management information 63. Each piece of volume management information 63 comprises: a write-sequence management information number 801 for specifying write-sequence management information 62 used by an own volume; copy-to information 802 of the volume in question; copy-from information 803 of the volume in question; a bit-map switching flag 804 used to exclude bit-map switching; the latest write sequence number 805 used when the volume in question is a copy-to volume; a bit-map switching trigger write sequence number 806 which is compared with the latest write sequence number so as to trigger bit-map switching; a write-sequence-number check request flag 807 which is used to request at the time of bit-map switching to check whether or not the latest write sequence number 805 reaches the bit-



map switching trigger write sequence number 806 in the own storage system 2; a write-sequence-number reach flag 808 for indicating that the latest write sequence number 805 has reached the bit-map switching trigger write sequence number 806; a plurality of bit maps 809, each of which stores updated points of the volume in question; a plurality of bit-map validity flags 810, each of which indicates that a corresponding bit map is valid; a plurality of bit-map write start write sequence numbers 811 each indicating the write sequence number from which writing to a corresponding bit map is started; a plurality of bit-map updated data amount counters 812 each indicating the amount of updated data of a corresponding bit map; and a difference bit map 813 which is used when it becomes necessary to make contents of a volume coincide with those in the other sites as a result of suffering from disaster.

[0040]

The copy-to information 802 and the copy-from information 803 include: information that is required to communicate with the copy-from or copy-to storage system 2; and information for identifying a copy-from or copy-to volume.

[0041]

According to the embodiment of the present invention, for the sake of simplification, management information is

provided on a volume basis (the volume management information 63). However, this does not limit the scope of the present invention. The management information can be provided in an arbitrary unit that can be shared between the storage systems 2. Accordingly, it is possible to make a remote copy in the arbitrary unit that can be shared between the storage systems 2. The write processing 31 which is executed in each storage system 2 will be described with reference to Fig. 4.

[0042]

Here, a write requester and classification of write data, which are targeted by the write processing 31, will be outlined so that the description can be easily understood. The write requester is the host computer 1, or the storage system 2 that is a source of a remote copy.

[0043]

Depending on a write target volume, write data from the host computer 1 can be classified into: 1) write data having no remote-copy destination; and write data having a remote-copy destination. The write data having a remote-copy destination can be further classified into: 2) write data having only a synchronous remote-copy destination; 3) write data having only an asynchronous remote-copy destination; and 4) write data having both a synchronous remote-copy destination and an asynchronous remote-copy

destination.

[0044]

Depending on a write target volume, write data from the storage system 2 which is a remote-copy source can be classified into: write data that is written by a synchronous remote copy; and write data that is written by an asynchronous remote copy. As is the case with the above 1) through 4), the write data that is written by a synchronous remote copy can be classified into: 5) write data having no remote-copy destination; and write data having a remote-copy destination. The write data having a remote-copy destination can be further classified into: 6) write data having only a synchronous remote-copy destination; 7) write data having only an asynchronous remote-copy destination; and 8) write data having both a synchronous remote-copy destination and an asynchronous remote-copy destination. In addition, as is the case with the above 1), 4), the write data which is written by an asynchronous remote copy can be classified into: 9) write data having no remote-copy destination; and 10) write data having only an asynchronous remote-copy destination. The reason why 2) and 3) are excluded from the targets is that in the synchronous remote copy, a write completion notification is transferred to the host computer on the completion of a remote copy. Accordingly, even if, in the middle copy-to volume V2,

remote-copy data received by an asynchronous remote copy is remote-copied to a volume in another site by a synchronous remote copy, it is a meaningless and unrealistic matter.

[0045]

In the write processing 31, first of all, a write request is received (step S401), and then a judgment is made as to whether or not conditions that a write requester is the host computer 1, and that a write target volume is not a remote-copy source, hold true (step S402).

[0046]

If both of the conditions in the step S402 hold true, the above write request is a write request from the host computer 1 to a volume that does not relate to a remote copy (more specifically, write data is equivalent to the 1) described above). Accordingly, write data is stored in a normal area of the cache memory 5 (step S403), and then the write requester is notified of write completion (step S404). The storage controller 21 writes this write data to the storage device 22 in asynchronization with the write request from the host computer 1.

[0047]

If one or both of the conditions in the step S402 does not hold true, a check is made as to whether or not the write requester is the host computer 1 (step S405).

[0048]

If the condition in the step S405 holds true (more specifically, the write data is equivalent to any of 2), 3), 4) described above), the write-sequence management information entry 721 to be associated with the write data is created, and then a write sequence number is given to the write-sequence management information entry 721. To be more specific, with reference to the write-sequence management information number 801 of the volume management information 63 corresponding to a volume to which a write has been made from the host computer 1, the write-sequence management information entry 721 of the write-sequence management information 62 corresponding to the write-sequence management information number 801 is created. After that, a value of the write-sequence number counter 71 is set, and then the value of the write-sequence number counter 71 is added so that the write-sequence management information entry 721 is associated with the write data.

[0049]

The addition of the write sequence number is executed regardless of synchronization/asynchronization of the remote copy. This is because even if direct copy destinations of the volume in question include only synchronous remote-copy destinations, a further destination of these copy destinations may execute an asynchronous remote copy.

[0050]

If the condition in the step S405 does not hold true (more specifically, the write data is equivalent to any of 6) through 10) described above), the write-sequence management information entry 721 to be associated with the write data is created, and then the write-sequence management information 62 including a write sequence number, which has been received together with the write data, is stored (step S407).

[0051]

Next, the write data is stored in the cache memory 5. A method for storing the write data in the cache memory 5 depends on a write target volume. There are three cases as follows: A) a case where the write target volume is an asynchronous remote-copy destination (more specifically, equivalent to any of 9), 10) described above); B) a case where the write target volume is not an asynchronous remote-copy destination but an asynchronous remote-copy source (more specifically, equivalent to any of 4), 7), 8) described above); and C) a case other than the above A), B) (more specifically, equivalent to any of 2), 5), 6) described above). This is because in the case of A) it is necessary to temporarily store received write data, and then formally write the write data according to the write sequence number. This is also because in the case of B) if untransferred write data is written to an asynchronous

remote-copy destination, it is necessary to independently store this write data until old write data is transferred.

[0052]

First of all, a judgment is made as to whether or not a write target volume is an asynchronous remote-copy destination (step S408).

[0053]

If the condition in the step S408 holds true (more specifically, equivalent to A) described above), it is necessary to write the write data in question according to a write sequence number indicating the update order from the host computer 1. Accordingly, the write data is temporarily stored in the temporary area of the cache memory 5 (step S409). Then, the write-sequence management information entry 721 corresponding to the write data is registered in the normalization target write-sequence management information 74 as target data to be normalized (step S410). In Fig. 3, write-sequence management information entries 721\_e, 721\_f correspond to the target data to be normalized.

[0054]

Moreover, a judgment is made as to whether or not a write target volume is a copy source of an asynchronous remote copy (step S411). If the condition in the step S411 holds true (more specifically, equivalent to any of 3), 4), 7), 8), 10) described above), the write-sequence management

information entry 721 corresponding to the write data is registered in the transmission target write-sequence management information 73 as target data to be transferred (step S412). In Fig. 3, write-sequence management information entries 721\_b, 721\_c, 721\_d, 721\_e correspond to the target data to be transferred.

[0055]

If the condition in the step S408 holds true, a further judgment is made as to whether or not the write target volume is a copy source of an asynchronous remote copy (step S413).

[0056]

If the condition in the step S413 holds true (more specifically, equivalent to B) described above), write data that has not yet been transferred to an asynchronous remote-copy destination is stored in the temporary area of the cache memory 5, and then the write data is stored in the normal area of the cache memory 5 (step S414). In Fig. 3, the write-sequence management information entries 721\_b, 721\_c, 721\_d correspond to the write data. Further, the bit-map update processing 32 is executed (step S415). The bit-map update processing 32 will be described later. After that, the process proceeds to the step 411.

[0057]

If the condition in the step S413 does not hold true



(more specifically, equivalent to C) described above), the write data in question is neither a copy destination of an asynchronous remote copy nor a copy source of the asynchronous remote copy. Accordingly, the write data is stored in the normal area of the cache memory 5 (step S416). In Fig. 3, write-sequence management information entries 721\_a correspond to the write data. Further, the bit-map update processing 32 is executed (step S417). The bit-map update processing 32 will be described later. After that, the process proceeds to the step 418.

[0058]

Next, a judgment is made as to whether or not the write target volume is a copy source of a synchronous remote copy (step S418). If the condition in the step S418 holds true (more specifically, equivalent to any of 2), 4), 6), 8) described above), write data and contents of the write-sequence management information entry 721 are transferred to a copy destination of the synchronous remote copy (step S419). With the object of shortening the time taken after receiving a write request before a write completion notification is transferred to the write requester, the synchronous remote copy in question may also be executed in parallel with respect to a plurality of copy destinations.

[0059]

Next, a judgment is made as to whether the write-

sequence management information 721 corresponding to the write data in question is registered in neither the transmission target write-sequence management information 73 nor the normalization target write-sequence management information 74 (step S420). If the condition in the step S420 holds true (more specifically, equivalent to any of 2), 5), 6) described above), the write-sequence management information entry 721 corresponding to the write data in question is not necessary. Therefore, the write-sequence management information entry 721 is released (step S421), which corresponds to a write-sequence management information entry 721\_g in Fig. 3.

[0060]

Lastly, a write completion notification is transferred to the write requester (step S404), before the processing ends.

[0061]

The asynchronous transmission processing 33 which is executed in a storage system 2 having a copy-from volume of an asynchronous remote copy among the storage systems 2 will be described with reference to Fig. 5. This processing may also be executed for each write-sequence management information 62.

[0062]

First of all, a target to be transferred is selected

from among the write-sequence management information entries 721 registered in the transmission target write-sequence management information 73 (step S501).

[0063]

Next, the selected write-sequence management information entry 721 and write data corresponding to the write-sequence management information entry 721 in question are transferred to an asynchronous remote-copy destination (step S502).

[0064]

Next, the transferred write-sequence management information entry 721 is deleted from the transmission target write-sequence management information 73 (step S503).

[0065]

Next, a judgment is made as to whether or not the transferred write-sequence management information entry 721 is registered also in the normalization target write-sequence management information 74 (step S504). If the condition in the step S504 does not hold true, management as target data of the asynchronous remote copy becomes unnecessary. Accordingly, the transferred write-sequence management information entry 721 is released (step S505). Here, if the transferred write data has been stored in the temporary area of the cache memory 5, the temporary area in question is also released (step S506).

[0066]

The asynchronous normalization processing 41 which is executed in a storage system 2 having a copy-to volume of an asynchronous remote copy among the storage systems 2 will be described with reference to Fig. 6. This processing may also be executed for each write-sequence management information 62.

[0067]

First of all, with reference to the write-sequence management information entry 721 registered in the normalization target write-sequence management information 74, a range within which write sequence numbers successively continue is selected as a target to be normalized (step S601).

[0068]

Next, a judgment is made as to whether or not all write data which is the selected target to be normalized has already been normalized (step S602).

[0069]

If the condition in the step S602 holds true, the process proceeds to a step S606.

[0070]

If the condition in the step S602 does not hold true, among the pieces of write data which have been selected as the target to be normalized, a piece of write data

corresponding to the oldest write sequence number is transferred from the temporary area to the normal area in the cache memory 5, and then the piece of write data is normalized there (step S603).

[0071]

Next, the normalized write-sequence management information entry 721 is deleted from the normalization target write-sequence management information 74 (step S604). Next, the bit-map update processing 32 is executed (step S605). The bit-map update processing 32 will be described later. After that, the process returns to the step S602.

[0072]

Next, a judgment is made as to whether or not the normalized write-sequence management information entry 721 is registered also in the transmission target write-sequence management information 73 (step S606). If the condition in the step S606 does not hold true, management as target data of the asynchronous remote copy becomes unnecessary.

Accordingly, the normalized write-sequence management information entry 721 is released (step S607). In addition, the temporary area of the cache memory 5, in which the normalized write data has been stored, is released (step S608).

[0073]

The bit-map switching start processing 34 which is

executed in a storage system 2 having the top copy-from volume V1 among the storage systems 2 will be described with reference to Fig. 7. This processing may also be executed on a volume basis if a value of an updated data amount counter 812 of the bit map 809 whose bit-map validity flag 810 is in an ON state exceeds a fixed threshold value (if the number of bit maps 809 whose bit-map validity flag 810 is in an ON state is two or more, a bit map whose write start write sequence number 811 is the oldest is selected from among them).

[0074]

In addition, if the elapsed time after the lastly executed bit map switching exceeds a fixed threshold value, the bit-map switching start processing 34 may also be executed on a volume basis.

[0075]

First of all, for the top copy-from volume V1, all middle copy-to volumes V2 corresponding to the top copy-from volume V1, and all end copy-to volumes V3 corresponding to the top copy-from volume V1, in order to prevent bit map switching from being newly started before the last bit map switching ends, a judgment is made as to whether or not the bit-map switching flag 804 is in an OFF state (step S701).

[0076]

If the condition in the step S701 does not hold true,

the bit map switching is still being executed. Therefore, the processing is ended.

[0077]

If the condition in the step S701 holds true, the bit-map switching flag 804 is brought into an ON state (step S702).

[0078]

Next, a write sequence number used to trigger bit-map switching is determined. Then, the bit-map switching trigger write sequence number 806 of the management information memory is set at the determined write sequence number (step S703). A method for determining a write sequence number used to trigger bit-map switching may be the following: A write sequence number is so generated that the latest write sequence number 805 relating to the middle copy-to volume V2 and the end copy-to volume V3 does not exceed the generated write sequence number before a notification is transferred from the top copy-from volume V1 to all middle copy-to volumes V2 corresponding to the top copy-from volume V1, and further to all end copy-to volumes V3 corresponding to the top copy-from volume V1. This write sequence number is used as the bit-map switching trigger write sequence number 806.

[0079]

In addition, it is intended that the bit-map

switching trigger write sequence number 806 becomes a write start write sequence number of a bit map 809 to which bit-map switching is made. However, if write sequences of a plurality of volumes are managed with a piece of write-sequence management information 62, the determined bit-map switching trigger write sequence number 806 is not always assigned to write data of the volume in question. However, a write sequence number that is first assigned to the volume in question with the determined bit-map switching trigger write sequence number 806 being exceeded is uniquely determined for the top copy-from volume V1, all middle copy-to volumes V2 corresponding to the top copy-from volume V1, and all end copy-to volumes V3 corresponding to the top copy-from volume V1. Therefore, the write sequence number that is first assigned to the volume in question with the bit-map switching trigger write sequence number 806 being exceeded can be used as a write start write sequence number of the bit map 809 to which bit-map switching is made.

[0080]

Next, the write-sequence-number reach flag 808 is brought into an OFF state (step S704). This write-sequence-number reach flag 808 is used to indicate that the latest write sequence number 805 becomes larger than or equal to the bit-map switching trigger write sequence number 806.

[0081]



Next, the write-sequence-number check request flag 807 is brought into an ON state (step S705). This write-sequence-number check request flag 807 is used to request the bit-map update processing 32 to compare the latest write sequence number 805 with the bit-map switching trigger write sequence number 806.

[0082]

Next, the bit-map switching trigger write sequence number 806 is transferred to all direct copy destinations of the volume in question to request a write-sequence-number reach notification when the latest write sequence number exceeds the bit-map switching trigger write sequence number 806 (step S706).

[0083]

Next, judgments are made as to whether or not the write-sequence-number reach flag 808 of the volume in question is ON, and whether or not write-sequence-number reach notifications are received from all of the direct copy destinations of the volume in question (step S707). If one or both of the conditions in the step S707 does not hold true, after waiting for a fixed period of time (step S708), the judgments in the step S707 are made again. It is to be noted that notifications from the direct copy destinations of the volume in question are successively stored in the copy-to information 802.

[0084]

If both of the conditions in the step S707 hold true, a validity flag 810 of a bit map 809 corresponding to the oldest write start write sequence number 811 is brought into an OFF state, said bit map 809 being selected from among the bit maps 809 in the own site. Then, the bit map 809 in question is cleared (step S709). This is because at this point of time, for the top copy-from volume V1, all middle copy-to volumes V2 corresponding to the top copy-from volume V1, and all end copy-to volumes V3 corresponding to the top copy-from volume V1, the latest write sequence numbers 805 of these volumes exceed the bit-map switching trigger write sequence number 806, and consequently writing to a new bit map 809 has already been started.

[0085]

Next, all of the direct copy destinations of the volume in question are requested to perform bit map switching (step S710).

[0086]

Next, a judgment is made as to whether or not bit map switching completion notifications are received from all of the direct copy destinations of the volume in question (step S711). If the condition in the step S711 does not hold true, after waiting for a fixed period of time (step S712), the judgment in the step S711 is made again.

[0087]

If the condition in the step S711 holds true, the bit map switching is completed. Therefore, the bit-map switching flag 804 is brought into an OFF state (step S713), before the processing ends.

[0088]

The bit-map switching processing 35 which is executed in a storage system 2 having the middle copy-to volume V2 and the end copy-to volume V3 among the storage systems 2 will be described with reference to Fig. 8.

[0089]

First of all, the bit-map switching trigger write sequence number 806 is received from a copy-from site. Then, the bit-map switching trigger write sequence number 806 of the middle copy-to volume V2, and that of the end copy-to volume V3, are set at a value of the received bit-map switching trigger write sequence number 806 (step S801). Here, the middle copy-to volume V2 and the end copy-to volume V3 are targets of bit map switching.

[0090]

Next, the write-sequence-number reach flag 808 is brought into an OFF state (step S802), and the write-sequence-number check request flag 807 is brought into an ON state (step S803).

[0091]

Next, a judgment is made as to whether or not a direct copy destination exists in the volume in question (step S804).

[0092]

If the condition in the step S804 does not hold true (more specifically, the volume in question is equivalent to the end copy-to volume V3), the process proceeds to a step S808.

[0093]

If the condition in the step S804 holds true (more specifically, the volume in question is equivalent to the middle copy-to volume V2), the bit-map switching trigger write sequence number 806 is transferred to all direct copy destinations of the volume in question to request a write-sequence-number reach notification when the latest write sequence number 805 exceeds the bit-map switching trigger write sequence number 806 (step S805).

[0094]

Next, a judgment is made as to whether or not write-sequence-number reach notifications are received from all of the direct copy destinations of the volume in question (step S806). If the condition in the step S806 does not hold true, after waiting for a fixed period of time (step S807), the judgment in the step S806 is made again.

[0095]

If the condition in the step S806 holds true, a judgment is made as to whether or not the write-sequence-number reach flag of the volume in question is ON (step S808). If the condition in the step S808 does not hold true, after waiting for a fixed period of time (step S809), the judgment in the step S808 is made again.

[0096]

If the condition in the step S808 holds true, the direct copy source of the volume in question is notified that the write sequence number has been reached (step S810).

[0097]

Next, a judgment is made as to whether or not a bit-map switching request has been received from a direct copy source of the volume in question (step S811). If the condition in the step S811 does not hold true, after waiting for a fixed period of time (step S812), the judgment in the step S811 is made again.

[0098]

If the condition in the step S811 holds true, the bit-map validity flag 810 of a bit map 809 corresponding to the oldest write start write sequence number 811 is brought into an OFF state, the bit map 809 being selected from among the bit maps 809 of the volume in question. Then, the bit map 809 in question is cleared (step S813).

[0099]

Next, a judgment is made as to whether or not a direct copy destination exists in the volume in question (step S814).

[0100]

If the condition in the step S814 does not hold true (more specifically, the volume in question is equivalent to the end copy-to volume V3), the process proceeds to a step S818.

[0101]

If the condition in the step S814 holds true (more specifically, the volume in question is equivalent to the middle copy-to volume V2), all of the direct copy destinations of the volume in question are requested to perform bit-map switching (step S815).

[0102]

Next, a judgment is made as to whether or not bit map switching completion notifications are received from all of the direct copy destinations of the volume in question (step S816). If the condition in the step S816 does not hold true, after waiting for a fixed period of time (step S817), the judgment in the step S816 is made again.

[0103]

If the condition in the step S816 holds true, the direct copy source of the volume in question is notified that the bit map switching has been completed (step S818),

and then the processing ends.

[0104]

The bit-map update processing 32, which is executed in the write processing 31 and the asynchronous normalization processing 41, will be described with reference to Fig. 9.

[0105]

First of all, each bit of a bit map 809 whose bit-map validity flag 810 is ON is brought into an ON state, the said bit corresponding to write data, and then the updated data amount counter 812 corresponding to the updated bit map 809 is incremented (step S901).

[0106]

Next, the latest write sequence number 805 is updated to a write sequence number corresponding to the write data in question (step S902).

[0107]

Next, a judgment is made as to whether or not the write-sequence-number check request flag 807 is ON (step S903).

[0108]

If the condition in the step S903 does not hold true, the processing ends.

[0109]

If the condition in the step S903 holds true, a

judgment is made as to whether or not the latest write sequence number 805 is larger than or equal to the bit-map switching trigger write sequence number 806 (step S904).

[0110]

If the condition in the step S904 does not hold true, the processing ends.

[0111]

If the condition in the step S904 holds true, the write-sequence-number reach flag 808 is brought into an ON state (step S905).

[0112]

Next, the bit-map validity flag 810 of a bit map 809 whose bit-map validity flag 810 is OFF is brought into an ON state, and then the write start write sequence number 811 is set at a value of the latest write sequence number 805 (step S906).

[0113]

Next, each bit of the bit map 809 whose bit-map validity flag 810 has been newly brought into the ON state is switched to ON, the said bit corresponding to write data, and then the updated data amount counter 812 corresponding to the updated bit map 809 is incremented (step S907).

[0114]

Lastly, the write-sequence-number check request flag 807 is brought into an OFF state (step S908), before ending



the processing.

[0115]

Fig. 10 is a diagram schematically illustrating processing in each site, and communications between sites, at the time of bit-map switching.

[0116]

The bit-map switching start processing 34 is executed in a site having the top copy-from volume V1; and the bit-map switching processing 35 is executed in a site having the middle copy-to volume V2 or the end copy-to volume V3. However, the bit-map update processing 32 is partially included. An appropriate step number is given to each step to be executed.

[0117]

Fig. 10 schematically illustrates bit-map switching from a bit map 1 to a bit map 2 in the whole site. After that, the bit-map switching is alternately executed from the bit map 2 to the bit map 1, and further from the bit map 1 to the bit map 2.

[0118]

By shifting the bit-map switching as the whole site from a state in which a single bit map is updated, through a state in which double bit maps are temporarily updated, to a state in which a single bit map is updated again, the write start write sequence number 811 can always share the same

bit map 809 between arbitrary sites during an arbitrary moment including the bit-map switching, and thereby volume contents can be made to coincide with those in the site in which the write sequence number goes ahead.

[0119]

In addition, in the event that a failure is detected in a certain site during the bit-map switching, transfer of the bit-map switching trigger write sequence number 806 to the storage system 2 in each site may fail. In this case, in the storage system 2 having the top copy-from volume V1, the bit map 809 is not cleared in the step S709. Then, in a step S710, by requesting storage systems 2 in the other sites to transfer a bit-map switching completion notification without clearing the bit map 809, it is possible to continue management of differences including the site where the failure has occurred.

[0120]

Moreover, if each storage system 2 has triple bit maps 809 and a failure is detected in a certain site, as described above, by preventing the bit map 809 used when the failure has been detected from being cleared, and by continuing the bit-map switching by use of remaining bit maps, it is possible to continue the management of the differences including the site where the failure has occurred, and thereby to prevent the differences between

normal sites from increasing. Even if the number of sites in which a failure has occurred increases, by keeping unchanged the bit map 809, which is prevented from being cleared, until all sites become normal, it is possible to continue the management of the differences including the site where the failure has occurred.

[0121]

In addition, by transferring and receiving usual remote-copy data together with the contents of communications between two storage systems performed at the time of the bit-map switching, it is possible to reduce a communication load between the sites.

[0122]

Steps of making the volume contents coincide between sites after one of the sites suffers from disaster will be described with reference to Fig. 11.

[0123]

First of all, as a new direct copy source of an own volume that has lost a direct copy source as a result of the disaster from which the site has suffered, a volume whose latest write sequence number 805 is larger than that of the own volume is selected from among volumes having the same top copy-from volume V1 as that included in the own volume (step S1101).

[0124]

Next, a transfer path is provided between the storage system 2 in an own site and the storage system 2 in a new direct copy-from site (step S1102).

[0125]

Next, remote copying is started between the storage system 2 in the own site and the storage system 2 in the new direct copy-from site (step S1103). Thereafter, updates made from the new direct copy-from site to the copy-from volume are successively reflected in a copy-to volume in the own site.

[0126]

Lastly, the difference copy processing 36 between the storage system 2 in the own site and the storage system 2 in the new direct copy-from site is started (step S1104). The difference copy processing 36 will be described later.

[0127]

By continuing remote copying also after completion of the difference copy processing 36, the remote copying becomes usual remote copy operation.

[0128]

As the new direct copy-from site described above, a site on the upstream side whose distance to the top copy-from volume V1 is closer than the distance from the own site to the top copy-from volume V1 may also be registered beforehand for the own volume.

[0129]

The difference copy processing 36 which is one of the steps of making the volume contents coincide between sites after one of the sites suffers from disaster will be described with reference to Fig. 12.

[0130]

First of all, a bit map 809 whose bit-map validity flag 810 is ON in both an own site and a new direct copy-from site is selected in the new direct copy-from site. Here, if there are a plurality of bit maps each satisfying this condition, a bit map 809 whose write start write sequence number 811 is the newest is selected from among the bit maps. Then, contents of the selected bit map 809 is copied to the difference bit map 813 that is used to determine data to be copied in the difference copy processing 36 (step S1201). The reason why the bit map 809 is copied to the difference bit map 813 is the following: In the new direct copy-from site, the remote copying may be continued even after the site has suffered from disaster, and as a result, the bit map 809 whose bit-map validity flag 810 is ON may be continuously updated. This makes it necessary to save the differences between current volumes in both sites so as to prevent the amount of differences to be copied from increasing.

[0131]

Next, part corresponding to bits whose value is ON in the difference bit map 813 is copied from a copy-from volume in the new direct copy-from site to a copy-to volume in the own site (step S1202).

[0132]

Lastly, the difference bit map 813 is cleared (step S1203), and then the processing ends.

[0133]

[Effects of the Invention]

According to the present invention, in remote copies made among storage controllers that are located in N sites (the number of sites is in general three or more), after one site suffers from disaster, it is possible to manage differences that are used to quickly make remote-copy data coincide with each other among the remaining sites. In addition, it is possible to keep the amount of information used to manage the differences constant without depending on the number of sites N among which remote copies are made.

[Brief Description of the Drawings]

[Fig. 1]

Fig. 1 is a diagram illustrating how remote copies among N sites are made according to an embodiment of the present invention by way of example.

[Fig. 2]

Fig. 2 is a diagram illustrating a configuration of a

storage system according to an embodiment of the present invention by way of example.

[Fig. 3]

Fig. 3 is a diagram illustrating the relationship between write-sequence management information and a cache memory according to an embodiment of the present invention by way of example.

[Fig. 4]

Fig. 4 is a flowchart illustrating an example of write processing that is executed in a storage system according to an embodiment of the present invention.

[Fig. 5]

Fig. 5 is a flowchart illustrating an example of asynchronous transmission processing that is executed in a storage system of an asynchronous remote copy source according to an embodiment of the present invention.

[Fig. 6]

Fig. 6 is a flowchart illustrating an example of asynchronous normalization processing that is executed in a storage system of an asynchronous remote copy destination according to an embodiment of the present invention.

[Fig. 7]

Fig. 7 is a flowchart illustrating an example of bit-map switching start processing that is executed in a storage system having a top copy-from volume V1 according to an

embodiment of the present invention.

[Fig. 8]

Fig. 8 is a flowchart illustrating an example of bit-map switching processing that is executed in a storage system having a middle copy-from volume/an end copy-from volume according to an embodiment of the present invention.

[Fig. 9]

Fig. 9 is a flowchart illustrating an example of bit-map update processing which is executed in write processing and asynchronous normalization processing according to an embodiment of the present invention.

[Fig. 10]

Fig. 10 is a diagram schematically illustrating bit-map switching according to an embodiment of the present invention.

[Fig. 11]

Fig. 11 is a flowchart illustrating steps of making the volume contents coincide between sites after one of the sites suffers from disaster according to an embodiment of the present invention by way of example.

[Fig. 12]

Fig. 12 is a flowchart illustrating an example of difference copy processing which is one of the steps of making the volume contents coincide between sites after one of the sites suffers from disaster according to an



embodiment of the present invention.

[Explanations of Symbols and Reference Numerals]

1 ... Host computer, 2 ... Storage system, 21 ... Storage controller, 22 ... Storage device, 23 ... Maintenance terminal, 3 ... Channel interface, 31 ... Write processing, 32 ... Bit-map update processing, 33 ... Asynchronous transmission processing, 34 ... Bit-map switching start processing, 35 ... Bit-map switching processing, 36 ... Difference copy processing, 4 ... Disk interface, 41 ... Asynchronous normalization processing, 5 ... Cache memory, 6 ... Management information memory, 62 ... Write-sequence management information, 63 ... Volume management information, 71 ... Write-sequence number counter, 721 ... Write-sequence management information entry, 73 ... Transmission target write-sequence management information, 74 ... Normalization target write-sequence management information, 804 ... Bit-map switching flag, 805 ... Latest write sequence number, 806 ... Bit-map switching trigger write sequence number, 807 ... Write-sequence-number check request flag, 808 ... Write-sequence-number reach flag, 809 ... Bit map, 810 ... Bit-map validity flag, 811 ... Bit-map write start write sequence number, 812 ... Bit-map updated data amount counter, 813 ... Difference bit map, T1 ... Global network, T2 ... Local network, V1 ... Top copy-from volume, V2 ... Middle copy-to volume, V3 ... End

copy-to volume.

Fig. 1

←Remote-copy data

Global network T1

Site 1

Host computer 1\_1

Local network T2\_2

Storage system 2\_1

Top copy-from volume V1\_1

Site 2

Host computer 1

Local network T2\_2

Storage system 2\_2

Middle copy-to volume V2\_2

Site (K-1)

Host computer 1\_(K-1)

Local network T2\_(K-1)

Storage system 2\_(K-1)

End copy-to volume V3\_(K-1)

Site (N-1)

Host computer 1\_(N-1)

Local network T2\_(N-1)

Storage system 2\_(N-1)  
 End copy-to volume V3\_(N-1)

Site (N-2)  
 Host computer 1\_(N-2)  
 Local network T2\_(N-2)  
 Storage system 2\_(N-2)  
 End copy-to volume V3\_(N-2)

Site K  
 Host computer 1\_K  
 Local network T2\_K  
 Storage system 2\_K  
 Middle copy-to volume V2\_K

Fig. 2

2: Storage system

21: Storage controller

3: Channel interface

31: Write processing

31: Write processing

32: Bit-map update processing

32: Bit-map update processing

33: Asynchronous transmission processing

33: Asynchronous transmission processing

34: Bit-map switching start processing

34: Bit-map switching start processing

35: Bit-map switching processing

35: Bit-map switching processing

36: Difference copy processing

36: Difference copy processing

4: Disk interface

41: Asynchronous normalization processing

41: Asynchronous normalization processing

32: Bit-map update processing

32: Bit-map update processing

22: Storage device

23: Maintenance terminal

5: Cache memory

Normal area

Temporary area

6: Management information memory

61: Remote copy management information

62\_1: Write-sequence management information 1

71: Write-sequence number counter  
72: Write-sequence management information entry area  
73: Transmission target write-sequence management  
information  
74: Normalization target write-sequence management  
information  
  
62\_(J-1): Write-sequence management information (J - 1)  
  
63\_1: Volume\_1 management information  
801: Write-sequence management information number  
802: Copy-to information  
803: Copy-from information  
804: Bit-map switching flag  
805: Latest write sequence number  
806: Bit-map switching trigger write sequence number  
807: Write-sequence-number check request flag  
808: Write-sequence-number reach flag  
809\_1: Bit map\_1  
810\_1: Bit-map\_1 validity flag  
811\_1: Bit-map\_1 write start write sequence number  
812\_1: Bit-map\_1 updated data amount counter  
809\_2: Bit map\_2  
810\_2: Bit-map\_2 validity flag  
811\_2: Bit-map\_2 write start write sequence number

812\_2: Bit-map\_2 updated data amount counter

813: Difference bit map

63\_(M-1) Volume number (M-1) management information

Fig. 3

2: Storage system

5: Cache memory

Normal area

Write data D0      Write data D1      Write data D2\_new

Temporary area

Write data D2\_old      Write data D4      Write data D5

6: Management information memory

61: Remote copy management information

62\_1: Write-sequence management information 1

71: Write-sequence number counter

72: Write-sequence management information entry area

721\_a: Write-sequence management information entry (write-sequence number)

721\_b Write-sequence management information entry (write-sequence number)

721\_c Write-sequence management information entry (write-

sequence number)

721\_d Write-sequence management information entry (write-sequence number)

721\_e Write-sequence management information entry (write-sequence number)

721\_f Write-sequence management information entry (write-sequence number)

721\_g Write-sequence management information entry (no entry)

73: Transmission target write-sequence management information

74: Normalization target write-sequence management information

62\_(J-1): Write-sequence management information (J-1)

Fig. 4

Start

S401: Receive a write request.

S402: Is it judged that a write requester is a host computer and a write target volume is not a remote-copy source?

S403: Store write data in a normal area of a cache memory.

S405: Is the write requester the host computer?

S406: Create a write-sequence management information entry to be associated with the write data, and give a write



sequence number to the write-sequence management information entry.

S407: Create a write-sequence management information entry to be associated with the write data, and store received write-sequence management information.

S408: Is the write target volume an asynchronous remote-copy destination?

S413: Is the write target volume an asynchronous remote-copy source?

S416: Store the write data in the normal area of the cache memory.

S417: Perform bit-map update processing.

S414: Store write data, which has not yet been transferred to an asynchronous remote-copy destination, in a temporary area of the cache memory, and then store the write data in the normal area of the cache memory.

S415: Perform bit-map update processing.

S409: Store the write data in the temporary area of the cache memory.

S410: Register the write-sequence management information entry as a target to be normalized.

S411: Is the write target volume an asynchronous remote-copy source?

S412: Register the write-sequence management information entry as a target to be transferred.

S418: Is the write target volume a synchronous remote-copy source?

S419: Transfer the write data and the write-sequence management information to a copy destination.

S420: Is it judged that the write-sequence management information entry is neither a target to be transferred nor a target to be normalized?

S421: Release the write-sequence management information entry.

S404: Notify the write requester of the write completion.

End

Fig. 5

Start

S501: Select a target to be transferred from among write-sequence management information entries registered as targets to be transferred.

S502: Transfer write data corresponding to a write-sequence management information entry to an asynchronous remote-copy destination.

S503: Delete the transferred write-sequence management information entry that has been registered as the target to be transferred.

S504: Is the transferred write-sequence management

information entry a target to be normalized?

S505: Release the transferred write-sequence management information entry.

S506: If a temporary area of a cache memory has been used by the transferred write data, release the temporary area.

End

Fig. 6

Start

S601: With reference to a write-sequence management information entry registered as a target to be normalized, select, as a target to be normalized, write data corresponding to write-sequence management information entries in a range within which write sequence numbers successively continue.

S602: Have all pieces of write data that have been selected as a target to be normalized been normalized?

S603: Among the pieces of write data that have been selected as the target to be normalized, transfer a piece of write data corresponding to the oldest write sequence number from the temporary area to the normal area in the cache memory.

S604: Delete the normalized write-sequence management information entry that has been registered as the target to be normalized.

S605: Perform bit-map update processing.  
S606: Is the normalized write-sequence management information entry a target to be transferred?  
S607: Release the normalized write-sequence management information entry.  
S608: Release a cache memory's temporary area that has been used by the normalized write data.  
End

Fig. 7

Start

S701: Is a bit-map switching flag in an OFF state?  
S702: Bring the bit-map switching flag into an ON state.  
S703: Determine a write sequence number used to trigger bit-map switching, and store the write sequence number in a management information memory.  
S704: Bring a write-sequence-number reach flag into an OFF state.  
S705: Bring a write-sequence-number check request flag into an ON state.  
S706: Transfer a bit-map switching trigger write sequence number to a direct copy destination of an own volume.  
S707: Is it judged that a write-sequence-number reach flag of the own volume is ON, and that write-sequence-number

reach notifications have been received from all of direct copy destinations of the own volume?

S708: Wait for a fixed period of time.

S709: Bring into an OFF state a validity flag of a bit map corresponding to the oldest write start write sequence number, and then clear the bit map, the bit map being selected from among bit maps of the own volume.

S710: Request the direct copy destinations of the own volume to perform bit-map switching.

S711: Have bit-map switching completion notifications been received from all of the direct copy destinations of the own volume?

S712: Wait for a fixed period of time.

S713: Bring the bit-map switching flag into an OFF state.

End

Fig. 8

Start

S801: Receive a bit-map switching trigger write sequence number from a direct copy source of an own volume, and store the bit-map switching trigger write sequence number in a management information memory.

S802: Bring a write-sequence-number reach flag into an OFF state.

S803: Bring a write-sequence-number check request flag into an ON state.

S804: Does a direct copy destination exist in the own volume?

S805: Transfer a bit-map switching trigger write sequence number 806 to a direct copy destination of the own volume.

S806: Have write-sequence-number reach notifications been received from all of direct copy destinations of the own volume?

S807: Wait for a fixed period of time.

S808: Is a write-sequence-number reach flag of the own volume in an ON state?

S809: Wait for a fixed period of time.

S810: Notify the direct copy source of the own volume that the write sequence number has been reached.

S811: Has a bit-map switching request been received from the direct copy source of the own volume?

S812: Wait for a fixed period of time.

S813: Bring into an OFF state a validity flag of a bit map corresponding to the oldest write start write sequence number, and then clear the bit map, the bit map being selected from among bit maps of the own volume.

S814: Does a direct copy destination exist in the own volume?

S815: Request the direct copy destinations of the own volume

to perform bit-map switching.

S816: Have bit-map switching completion notifications been received from all of the direct copy destinations of the own volume?

S817: Wait for a fixed period of time.

S818: Notify the direct copy source of the own volume that the bit-map switching has been completed.

End

Fig. 9

Start

S901: Bring into an ON state each bit of a bit map whose validity flag is ON, the bit corresponding to write data, and then increment a value of an updated data amount counter.

S902: Set the latest write sequence number at a write sequence number corresponding to the write data in question.

S903: Is a write-sequence-number check request flag in an ON state?

S904: Is the latest write sequence number larger than or equal to the bit-map switching trigger write sequence number?

S905: Bring the write-sequence-number reach flag into an ON state.

S906: Bring into an ON state a validity flag of a bit map

whose validity flag is OFF, and then set a write start write sequence number at a value of the latest write sequence number.

S907: Bring into an ON state each bit of a bit map whose validity flag has been newly brought into an ON state, the bit corresponding to write data, and then increment a value of the updated data amount counter.

S908: Bring the write-sequence-number check request flag in an OFF state.

End

Fig. 10

V1: Top copy-from volume

V2: Middle copy-to volume

V3: End copy-to volume

Bit map 1      Bit map 2

Detect the timing in which the bit-map switching is triggered.

S702: Bring the bit-map switching flag into an ON state.

A state in which a validity flag is ON

A state in which a validity flag is OFF



S706: Transfer a bit-map switching trigger write sequence number.

S904, S906, S907: When a bit map is updated, the latest write sequence number M is reached ( $\geq N$ ), and start an update to a bit map whose validity flag is OFF with a write start write sequence number.

S805: Transfer a bit-map switching trigger write sequence number N.

S904, S906, S907 When a bit map is updated, the latest write sequence number M is reached ( $\geq N$ ), and start an update to a bit map whose validity flag is OFF with a write start write sequence number.

S904, S906, S907: When a bit map is updated, the latest write sequence number M is reached ( $\geq N$ ), and start an update to a bit map whose validity flag is OFF with a write start write sequence number.

S808: Check whether or not the write sequence number N has been reached.

S810: Notify that the write sequence number N has been reached.

S808: Check whether or not the write sequence number N has been reached.

S810: Notify that the write sequence number N has been reached.

S707: Check whether or not the write sequence number N has

been reached.

S709: Bring into an OFF state a validity flag of a bit map corresponding to the oldest write start write sequence number, and then clear the contents of the bit map.

S710: Request to perform bit-map switching.

S813: Bring into an OFF state a validity flag of a bit map corresponding to the oldest write start write sequence number, and then clear the contents of the bit map.

S815: Request to perform bit-map switching.

S813: Bring into an OFF state a validity flag of a bit map corresponding to the oldest write start write sequence number, and then clear the contents of the bit map.

S818: Notify that the bit-map switching has been completed.

S818: Notify that the bit-map switching has been completed.

S713: Bring the bit-map switching flag into an OFF state.

Fig. 11

Start

S1101: As a new direct copy source of an own volume, select a volume whose latest write sequence number is larger than that of the own volume from among volumes having the same top copy-from volume as that included in the own volume.

S1102: Provide a transmission path between a storage system in an own site and a storage system in a new direct copy-

from site.

S1103: Start remote copying between the storage system in the own site and the storage system in the new direct copy-from site.

S1104: Start difference copy processing between the storage system in the own site and the storage system in the new direct copy-from site.

End

Fig. 12

Start

S1201: Select, in a new direct copy-from site, a bit map whose validity flag is ON in both an own site and the new direct copy-from site, and if there are a plurality of bit maps each satisfying this condition, copy a bit map whose write start write sequence number is the newest to create a difference bit map.

S1202: Copy an area corresponding to bits whose value is ON in the difference bit map from a copy source to a copy destination.

S1203: Clear the difference bit map.

End

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2003-131917

(P2003-131917A)

(43) 公開日 平成15年5月9日 (2003.5.9)

(51) Int.Cl. <sup>7</sup>	識別記号	F I	テマコード* (参考)
G 0 6 F 12/00	5 3 3	G 0 6 F 12/00	5 3 3 A 5 B 0 6 5
	5 4 5		5 4 5 A 5 B 0 8 2
3/06	3 0 4	3/06	3 0 4 F
			3 0 4 P

審査請求 未請求 請求項の数 8 O L (全 19 頁)

(21) 出願番号 特願2001-327112 (P2001-327112)

(22) 出願日 平成13年10月25日 (2001.10.25)

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 森下 昇

神奈川県川崎市麻生区王禅寺1099番地 株

式会社日立製作所システム開発研究所内

(72) 発明者 中野 俊夫

神奈川県小田原市中里322番地2号 株式

会社日立製作所 R A I D システム事業部内

(74) 代理人 100075096

弁理士 作田 康夫

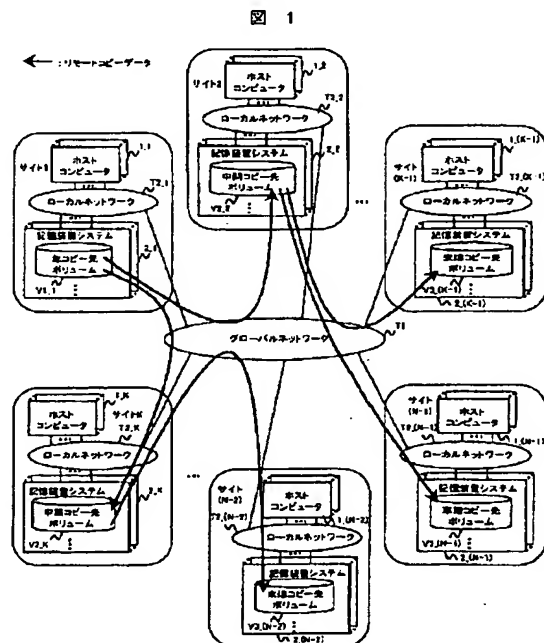
最終頁に続く

(54) 【発明の名称】 記憶装置システム

(57) 【要約】

【課題】主に3サイト以上のNサイトに配置した記憶制御装置間のリモートコピーにおいて、任意サイト被災後、速やかに残サイト間のリモートコピーデータの一致を実現するための差分を管理することを課題とする。

【解決手段】ホストコンピュータから直接更新されるリモートコピー範囲と、当該リモートコピー範囲のコピー先となる他サイトのリモートコピー範囲の更新履歴の記録を、当該サイト間で共通に認識した、ホストコンピュータからのある更新を契機に開始し、任意サイトの被災時、被災していない任意サイト間のリモートコピー範囲の内容を、更新履歴に従って、リモートコピー範囲の一部をコピーすることで、リモートコピー範囲の内容を一致させる。



【特許請求の範囲】

【請求項1】 ホストコンピュータと、ホストコンピュータと接続する記憶装置システムを備える複数のサイトにおいて、サイト間のネットワークにより記憶装置システム同士を接続し、記憶装置システムが、ホストコンピュータから更新されるデータを他サイトの記憶装置システムにも反映するリモートコピー構成において、ホストコンピュータから直接更新される記憶装置システムのリモートコピー指定範囲と、リモートコピー指定範囲のリモートコピー先となる他サイトの記憶装置システムのコピー先リモートコピー指定範囲の更新履歴の記録を、リモートコピー指定範囲、および、コピー先リモートコピー指定範囲を持つ記憶装置システム間で共通に認識した、ホストコンピュータからのある更新を契機に開始し、任意サイトの被災時、被災していない任意サイト間のリモートコピー指定範囲、または、コピー先リモートコピー指定範囲の内容を、更新履歴に従って、リモートコピー指定範囲、または、コピー先リモートコピー指定範囲の一部をコピーすることで、リモートコピー指定範囲、または、コピー先リモートコピー指定範囲の内容を一致させることを特徴とする記憶装置システム。

【請求項2】 請求項1に記載の記憶装置システムであって、リモートコピー指定範囲の更新履歴を記録する領域を多重に持ち、

ホストコンピュータから直接更新される記憶装置システムのリモートコピー指定範囲への、ホストコンピュータからの更新量が閾値を超えた契機において、新たな更新履歴への切り替え契機となるホストコンピュータからの更新を決定し、リモートコピー指定範囲、および、コピー先リモートコピー指定範囲を持つ記憶装置システム間で共有し、

各記憶装置システムにおいて、リモートコピー指定範囲、または、コピー先リモートコピー指定範囲の更新が、新たな更新履歴への切り替え契機となるホストコンピュータからの更新に到達した時点から、新たな更新履歴の記録を開始し、

リモートコピー指定範囲、および、コピー先リモートコピー指定範囲を持つ記憶装置システム全てが、新たな更新履歴の記録を開始したことを確認後、従来の更新履歴の記録を中断し、従来の更新履歴を初期化することを特徴とする、請求項1に記載の記憶装置システム。

【請求項3】 請求項2に記載の記憶装置システムであって、新たな更新履歴への切り替え契機となる、ホストコンピュータからの更新の決定を、前回の更新履歴の切り替えからの経過時間が閾値を超えた契機にも実行することを特徴とする、請求項2に記載の記憶装置システム。

【請求項4】 請求項2に記載の記憶装置システムであって、記憶装置システム間での情報のやりとりを、当該記憶装置システムのリモートコピー指定範囲、または、

コピー先リモートコピー指定範囲に対応する直接のコピー元、および、直接のコピー先との間でのみ実行することを特徴とする、請求項2に記載の記憶装置システム。

【請求項5】 請求項2に記載の記憶装置システムであって、記憶装置システム間での情報のやりとりを、リモートコピーデータの転送、および、その応答に添付して実行することを特徴とする、請求項2に記載の記憶装置システム。

【請求項6】 請求項2に記載の記憶装置システムであって、任意サイトの被災時、被災していない任意サイト間のリモートコピー指定範囲、または、コピー先リモートコピー指定範囲の内容を一致させるための更新履歴を、

両サイトにおいて、有効、かつ、記録を開始したホストコンピュータからの更新が同じであるという条件で選択し、

さらに、両サイトにおいて、前記条件を充足する複数の更新履歴が存在する場合には、記録を開始したホストコンピュータからの更新が新しい更新履歴を選択することを特徴とする、請求項2に記載の記憶装置システム。

【請求項7】 請求項2に記載の記憶装置システムであって、任意サイトの被災を検出した場合、従来の更新履歴の記録を続行し、被災したサイトがすべて正常となった時点で、当該従来の更新履歴の記録を中断し、初期化することを特徴とする、請求項2に記載の記憶装置システム。

【請求項8】 請求項2に記載の記憶装置システムであって、任意サイトの被災を検出した場合、従来の更新履歴を除外した更新履歴の領域により、被災していないサイト間で請求項2に記載の手順を実行することを特徴とする、請求項2に記載の記憶装置システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、ホストコンピュータを介さず、記憶装置システム間でデータを2重化するリモートコピー機能に関する。特に、3サイト以上に配置した記憶装置システム間のリモートコピーにおいて、任意サイトが被災し、被災していないサイト間でリモートコピーを継続する場合における、リモートコピー指定範囲の内容の一致管理に関する。

【0002】

【従来の技術】災害等による記憶装置システム内のデータ喪失を避けるため、遠隔地にある記憶装置システムに、当該データを2重化する機能にリモートコピーがある。

【0003】リモートコピーでは、ホストコンピュータに接続している正記憶制御装置と、正記憶制御装置と接続している副記憶制御装置に、リモートコピー対象ボリュームを設定し、正記憶制御装置に接続している正記憶装置の正ボリュームと、副記憶制御装置に接続している副記憶装置の副ボリュームの内容を常に一致させるよう

にコピーを実行し続ける。

【0004】特開平11-85408号公報には、異なる記憶制御装置間で、ホストコンピュータを介さずデータを2重書きする技術が開示されている。本発明では、ホストコンピュータからライト時刻を付加されたライトデータを受領した正記憶制御装置は、ホストコンピュータにライトデータのライト完了報告後、ライト時刻順に、ライト時刻とライトデータを副記憶制御装置に転送する。副記憶制御装置では、正記憶制御装置から受領したライト時刻とライトデータを不揮発のキャッシュメモリに格納し、あるライト時刻までのデータを保証する。

【0005】リモートコピーの動作は大きく2つに分けられる。ホストコンピュータへのライト要求の終了報告前に副記憶制御装置にライトデータを転送する同期リモートコピー、ホストコンピュータへのライト要求の終了報告後にライト要求とは非同期に副記憶制御装置にライトデータを転送する非同期リモートコピーである。データベース等のアプリケーションにリモートコピーを使用しようとする、ログの更新順序を保証するために、正ボリュームに対する更新順序どうりに副ボリュームを更新する必要がある。これを実現するために、非同期リモートコピーでは、正記憶制御装置において、ホストコンピュータからのライト毎に、ライトデータにシーケンス#を付加し、副記憶制御装置でシーケンス#どうりに副ボリュームに反映する方法がある。

【0006】上記に述べた従来のリモートコピーにおいて、リモートコピーを一時中断する場合（例えば、正記憶制御装置と副記憶制御装置間のリンク障害等によってリモートコピーが中断する場合等）、リモートコピー再開時に正ボリュームと副ボリュームの内容を早く一致させることを目的に、リモートコピーが中断している間のホストコンピュータからの正ボリュームに対する更新履歴を正/副ボリューム間の差分情報として正記憶制御装置内に記録しておくことがある。その後、リモートコピー再開時に、記録しておいた更新履歴を使用して、正ボリュームの内容を副ボリュームへ反映することで、正ボリュームと副ボリュームの内容を一致させることができる。

【0007】従来のリモートコピーにおける差分情報は、通常、ホストコンピュータからの更新位置であって、正ボリュームは更新されているが副ボリュームは更新されていない位置を記録する性質のものであるため、正記憶制御装置側に持つのが一般的である。

【0008】近年、ネットワークの高速化、低価格化に伴い、遠隔地間のデータ転送コストが下がりつつある。このため、リモートコピーにおける転送データ量を増やし、リモートコピーデータの可用性をさらに高めるために、1サイトが被災してもデータの2重化を維持できるように、3サイト以上のNサイト間でリモートコピーを実施したいという要求がある。

【0009】Nサイト間リモートコピーの運用として、Nサイトの内1サイトが被災した場合、残りの(N-1)サイトでリモートコピーを再構成し、リモートコピーを継続する形態が望まれる。この場合、リモートコピーを継続するために、残りの(N-1)サイト間でリモートコピーデータを速やかに一致させる必要がある。つまり、(N-1)サイト間の差分管理を行い、差分相当のコピーによって速やかにリモートコピーデータを一致させることが望ましい。

【0010】なぜなら、差分管理を実施しない場合、全てのリモートコピーデータをコピー元からコピー先へコピーする必要が生じる場合がある。これは、差分相当のコピーと比較し、コピーに長時間かかる上、コピー実行中においてデータの可用性が低下した状態となるため、確保しているサイト数だけの冗長性を生かしきれない。

【0011】ここで、例として、3サイトでリモートコピーを構成し、従来のリモートコピーにおける差分管理の方法を使用した場合の問題点を考える。3つのサイトをA、B、Cとする。

【0012】まず、サイト間を転送するリモートコピーデータの流れがA→B→Cの場合、サイトBが被災すると、サイトAとサイトCのボリューム間で差分をコピーしリモートコピーを再開することが望ましい。従来のリモートコピーにおける差分管理情報は正記憶制御装置側に持つのが一般的であることから、サイトA-B間の差分管理情報はサイトAが、サイトB-C間の差分管理情報はサイトBが持つことになる。このため、サイトB被災時には、サイトB-C間の差分管理情報が失われるため、サイトA-C間の差分が分からなくなる。

【0013】また、サイト間を転送するリモートコピーデータの流れがA→B、A→Cの場合、サイトAが被災すると、サイトBとサイトCのボリューム間で差分をコピーしリモートコピーを再開することが望ましい。この場合、サイトA-B、A-C間の差分管理情報は被災したサイトAが持っているため、サイトB-C間の差分が分からなくなる。

【0014】この問題に対応するため、通常時にリモートコピーデータのやりとりの無いサイト間で、あらかじめ差分管理を実施しておく方法が考えられる。上記の例においては、サイトB被災前にサイトA-C間の差分を、また、サイトA被災前にサイトB-C間の差分を管理しておく方法である。しかし、この方法の場合、サイト数に比例してあらかじめ管理しなければならないサイト間の差分情報が増大する、また、例えば、2サイト同時に被災するような想定外の事態に対し差分管理できない。

【0015】

【発明が解決しようとする課題】従来の、2サイト間でのリモートコピーの差分管理方法は、Nサイト間での差分管理について考慮していない。

【0016】このため、Nサイト間でのリモートコピー時、1つのサイトが被災し、被災していない残りの(N-1)サイトでリモートコピーを継続しようとした場合、すべてのリモートコピーデータを、新たなコピー元サイトからコピー先サイトへ転送してからでなければ、リモートコピーを再構成できないという問題があった。

【0017】また、従来の、2サイト間でのリモートコピーの差分管理方法を、サイト被災後のリモートコピー再構成時の新たなサイト間ペアに対してあらかじめ適応しておく場合、サイト数が増加するに従い、各サイトの差分管理情報が増大するという問題があった。また、多重サイトの被災に対して差分管理できないという問題があった。

【0018】本発明の目的は、主に3サイト以上のNサイトに配置した記憶装置システム間のリモートコピーにおいて、任意サイトの被災後、速やかに残サイト間のリモートコピーデータの一致を実現するための差分を管理することにある。また、リモートコピーを構成するサイト数Nに依存せず、差分を管理する情報量を一定に保つことにある。

【0019】

【課題を解決するための手段】この目的を達成するために、本発明による記憶装置システムは、ホストコンピュータと、ホストコンピュータと接続する記憶装置システムを備える複数のサイトにおいて、サイト間のネットワークにより記憶装置システム同士を接続し、記憶装置システムが、ホストコンピュータから更新されるデータを他サイトの記憶装置システムにも反映するリモートコピー構成において、ホストコンピュータから直接更新される記憶装置システムのリモートコピー指定範囲と、リモートコピー指定範囲のリモートコピー先となる他サイトの記憶装置システムのコピー先リモートコピー指定範囲の更新履歴を、リモートコピー指定範囲、および、コピー先リモートコピー指定範囲を持つ記憶装置システム間で共通に認識した、ホストコンピュータからのある更新を契機に記録を開始し、任意サイトの被災時、被災していないサイト間のリモートコピー指定範囲、または、コピー先リモートコピー指定範囲の内容を、更新履歴に従って、リモートコピー指定範囲、または、コピー先リモートコピー指定範囲の一部をコピーすることで、リモートコピー指定範囲、または、コピー先リモートコピー指定範囲の内容を一致させても良い。

【0020】また、リモートコピー指定範囲の更新履歴を記録する領域を多重に持ち、ホストコンピュータから直接更新される記憶装置システムのリモートコピー指定範囲への、ホストコンピュータからの更新量が閾値を超えた契機において、新たな更新履歴への切り替え契機となるホストコンピュータからの更新を決定し、リモートコピー指定範囲、および、コピー先リモートコピー指定範囲を持つ記憶装置システム間で共有し、各記憶装置シ

ステムにおいて、リモートコピー指定範囲、または、コピー先リモートコピー指定範囲の更新が、新たな更新履歴への切り替え契機となるホストコンピュータからの更新に到達した時点から、新たな更新履歴の記録を開始し、リモートコピー指定範囲、および、コピー先リモートコピー指定範囲を持つ記憶装置システム全てが、新たな更新履歴の記録を開始したことを確認後、従来の更新履歴の記録を中断し、従来の更新履歴を初期化しても良い。

【0021】また、新たな更新履歴への切り替え契機となる、ホストコンピュータからの更新の決定を、前回の更新履歴の切り替えからの経過時間が閾値を超えた契機にも実行しても良い。

【0022】また、記憶装置システム間での情報のやりとりを、当該記憶装置システムのリモートコピー指定範囲、または、コピー先リモートコピー指定範囲に対応する直接のコピー元、および、直接のコピー先との間でのみ実行しても良い。

【0023】また、記憶装置システム間での情報のやりとりを、リモートコピーデータの転送、および、その応答に添付して実行しても良い。

【0024】また、任意サイトの被災時、被災していないサイト間のリモートコピー指定範囲、または、コピー先リモートコピー指定範囲の内容を一致させるための更新履歴を、両サイトにおいて、有効、かつ、記録を開始したホストコンピュータからの更新が同じであるという条件で選択し、さらに、両サイトにおいて、前記条件を充足する複数の更新履歴が存在する場合には、記録を開始したホストコンピュータからの更新が新しい更新履歴を選択しても良い。

【0025】また、保守端末より、リモートコピー指定範囲、または、コピー先リモートコピー指定範囲のリモートコピー先毎に、リモートコピー先障害時の他のリモートコピー先へのリモートコピーの実行可否を指定し、リモートコピー先障害時、前記指定に従って、当該記憶装置システムからのリモートコピー先へのリモートコピーを制御しても良い。

【0026】

【発明の実施の形態】以下、本発明の第1の実施の形態を図面を用いて説明する。

【0027】図1に、Nサイト間でのリモートコピー構成の示す。各サイトには、複数のホストコンピュータ1や複数の記憶装置システム2等をLAN(Local Area Network)やSAN(Storage Area Network)といったローカルネットワークT2によって互いに接続している。一方、各サイトの記憶装置システム2の内、リモートコピーを実行する記憶装置システム2同士はグローバルネットワークT1を介し接続している。グローバルネットワークT1は、一般的に公衆通信回線であり、通信サービスを提供

する業者から有料で借り受けるものを利用する場合が多い。ただし、ローカルネットワークT2、グローバルネットワークT1の構成が、本発明を限定するものではない。

【0028】リモートコピー対象ボリュームの種類は、コピー元ボリュームを持たずコピー先ボリュームを持つ総コピー元ボリュームV1、コピー元ボリュームとコピー先ボリュームを持つ中間コピー先ボリュームV2、コピー元ボリュームを持ちコピー先ボリュームを持たない末端コピー先ボリュームV3に大別できる。リモートコピー対象ボリューム間の関係は、総コピー元ボリュームV1を頂点とするツリー状となり、総コピー元ボリュームV1から中間コピー先ボリュームV2を経て末端コピー先ボリュームV3へ順にリモートコピーデータを転送する。

【0029】図1では、サイト1のボリュームを総コピー元ボリュームV1とし、サイト2の中間コピー先ボリュームV2を経て、サイト(K-1)、サイトNの末端コピー先ボリュームV3へ、また、サイトKの中間コピー先ボリュームV2を経て、サイト(N-1)の末端コピー先ボリュームV3へリモートコピーを実施している構成である。

【0030】図2に、記憶装置システム2の構成を示す。記憶制御装置21は、ホストコンピュータ1と接続するチャネルインタフェース3と、記憶装置22と接続するディスクインタフェース4と、管理情報を格納する2重化した不揮発の管理情報メモリ6と、データを格納する2重化した不揮発のキャッシュメモリ5とをバスで接続した構成である。各チャネルインタフェース3、および、各ディスクインタフェース4は、2重化した管理情報メモリ6それぞれ、および、2重化キャッシュメモリ5それぞれと接続している。また、記憶制御装置21は、当該記憶制御装置21への指示や当該記憶制御装置21内部の状態を表示するための保守端末23を備える。

【0031】チャネルインタフェース3は、ホストコンピュータ1からのライト要求やリモートコピー元となる他の記憶装置システム2からのリモートコピーのライト要求を処理するライト処理31、ライトされた範囲に対応するビットマップを更新するビットマップ更新処理32、自記憶装置システム2が非同期リモートコピーのコピー元である場合、ライトデータをライト要求とは非同期にコピー先である他の記憶装置システム2に転送する非同期転送処理33、総コピー元ボリュームV1を持つ記憶装置システム2から当該ボリュームのコピー先となる全ボリュームのビットマップ809を切り替える契機を与えるビットマップ切り替え起動処理34、ビットマップ切り替え起動処理34からのビットマップ切り替え要求を受領し、ビットマップを切り替えるビットマップ切り替え処理35、任意サイトの被災後において残サイ

ト間でリモートコピー構成を再構成する場合の残サイト間のボリューム内容を一致させる差分コピー処理36を備える。

【0032】ディスクインタフェース4は、自記憶装置システム2が非同期リモートコピーのコピー先である場合、コピー元からのライト時に仮データとして一旦保存しておいたリモートコピーデータを正式化する非同期正式化処理41、チャネルインタフェース3側と同じビットマップ更新処理32を備える。

【0033】各処理はそれぞれ複数存在し、存在する単位は、ビットマップ切り替え起動処理34、ビットマップ切り替え処理35、および、差分コピー処理36がボリューム単位であったり、非同期転送処理33、および、非同期正式化処理41が後述するライトシーケンス管理情報62単位であったりしても良い。

【0034】キャッシュメモリ5は、記憶装置22に格納しているホストコンピュータ1からリード/ライトされるデータを格納している。キャッシュメモリ5の構成は、通常領域と仮領域に大別できる。仮領域を持つ理由は、自記憶装置システム2が非同期リモートコピーのコピー元である場合、コピー先に未転送のリモートコピーデータが再びライトされた際にライト前の未転送データを保存するため、あるいは、自記憶装置システム2が非同期リモートコピーのコピー先である場合、コピー元より受領し非同期正式化処理41によって正式化するまでのリモートコピーデータを保存するためである。

【0035】管理情報メモリ6は、記憶装置システム2が動作するために必要な管理情報を格納している。その中には、リモートコピー管理情報61が含まれ、リモートコピー管理情報61には、ライトシーケンス管理情報62、ボリューム管理情報63が含まれる。

【0036】ライトシーケンス管理情報62は、非同期リモートコピーにおいてホストコンピュータ1からのライト順序でリモートコピーデータを正式化するために使用する。ライトシーケンス管理情報62は複数存在し、各々のライトシーケンス管理情報62は、ホストコンピュータ1からのライト順序を記録するためのライトシーケンス#カウンタ71、ライトデータ毎に割り当てライトシーケンス#を保存しライトデータに関連付けるライトシーケンス管理情報エントリ領域72のライトシーケンス管理情報エントリ721、非同期リモートコピーのコピー元においてコピー先に未転送のリモートコピーデータを登録する転送対象ライトシーケンス管理情報73、非同期リモートコピーのコピー先においてコピー元から受領し未だ正式化していないリモートコピーデータを登録する正式化対象ライトシーケンス管理情報74等から構成する。

【0037】図3に、ライトシーケンス管理情報62の構造とキャッシュメモリ5の関係を示す。ライトデータ毎にライトシーケンス管理情報エントリ721を割り当



て、ライトシーケンス管理情報エントリ 7 2 1 にキャッシュメモリ 5 のライトデータの位置を設定する。また、転送対象ライトシーケンス管理情報 7 3、正式化対象ライトシーケンス管理情報 7 4 は、ライトシーケンス管理情報エントリ 7 2 1 を接続するキュー構造であっても良い。

【0038】また、ホストコンピュータ 1 からのライト順序を複数のボリューム間で保証したい場合には、同じライトシーケンス管理情報 6 2 のライトシーケンス # カウンタ 7 1 を使用して複数のボリュームのライトデータにライトシーケンス # を割り当て、同じ転送対象ライトシーケンス管理情報 7 3 に登録し、コピー先において、同じ正式化対象ライトシーケンス管理情報 7 4 に登録し、ライトシーケンス # に従って正式化すれば良い。

【0039】図 2 に戻って、ボリューム管理情報 6 3 は、リモートコピー対象ボリュームに関する管理情報である。ボリューム管理情報 6 3 は複数存在し、各々のボリューム管理情報 6 3 は、自ボリュームが使用するライトシーケンス管理情報 6 2 を指定するライトシーケンス管理情報 # 8 0 1、当該ボリュームのコピー先情報 8 0 2、当該ボリュームのコピー元情報 8 0 3、ビットマップ切り替えを排他するためのビットマップ切り替え中フラグ 8 0 4、当該ボリュームがコピー先ボリュームである場合の最新ライトシーケンス # 8 0 5、最新ライトシーケンス # と比較しビットマップを切り替える契機とするビットマップ切り替え契機ライトシーケンス # 8 0 6、ビットマップ切り替え時に自記憶装置システム 2 において最新ライトシーケンス # 8 0 5 がビットマップ切り替え契機ライトシーケンス # 8 0 6 に達したかどうかの確認を要求するライトシーケンス # 確認要求フラグ 8 0 7、最新ライトシーケンス # 8 0 5 がビットマップ切り替え契機ライトシーケンス # 8 0 6 に達したことを示すライトシーケンス # 到達フラグ 8 0 8、当該ボリュームの更新箇所を記録するビットマップ 8 0 9 (複数)、対応するビットマップが有効であることを示すビットマップ有効フラグ 8 1 0 (複数)、対応するビットマップへの記録を開始したライトシーケンス # を示すビットマップ記録開始ライトシーケンス # 8 1 1 (複数)、対応するビットマップへの更新量を示すビットマップ更新データ量カウンタ 8 1 2 (複数)、サイト被災等により他サイトとのボリューム内容を一致させる必要が生じた場合に使用する差分ビットマップ 8 1 3 等から構成する。

【0040】コピー先情報 8 0 2、コピー元情報 8 0 3 とは、コピー元/先の記憶装置システム 2 と通信するために必要な情報や、コピー元/先のボリュームを特定する情報等である。

【0041】また、簡単のため、本発明の実施の形態においては、管理情報をボリューム単位に持つ (ボリューム管理情報 6 3) こととしたが、これは、本発明の範囲を制限するものではない。管理情報は、記憶装置システ

ム 2 間で共有できる任意の単位で持つことができ、よって、記憶装置システム 2 間で共有できる任意の単位でリモートコピーを実行することができる。図 4 に、各記憶装置システム 2 で実行するライト処理 3 1 について説明する。

【0042】説明を分かり易くするため、ライト処理 3 1 が対象とする、ライト要求元とライトデータの種別の概要を補足する。ライト要求元は、ホストコンピュータ 1、または、リモートコピー元の記憶装置システム 2 である。

【0043】ホストコンピュータ 1 からのライトデータは、ライト対象ボリュームにより、1) リモートコピー先の無いもの、リモートコピー先の有るもので 2) 同期リモートコピー先のみのもの、3) 非同期リモートコピー先のみのもの、4) 両方のものに分けられる。

【0044】リモートコピー元の記憶装置システム 2 からのライトデータは、ライト対象ボリュームにより、同期リモートコピーによりライトされたもの、非同期リモートコピーによりライトされたものに分けられ、同期リモートコピーによりライトされたものは、上記 1) ~ 4) と同様に、5) リモートコピー先の無いもの、リモートコピー先の有るもので 6) 同期リモートコピー先のみのもの、7) 非同期リモートコピー先のみのもの、8) 両方のものに分けられる。また、非同期リモートコピーによるものは、上記 1)、4) と同様に、9) リモートコピー先の無いもの、10) 非同期リモートコピー先のみのものに分けられる。2)、3) を対象外としたのは、同期リモートコピーがリモートコピー完了後にホストコンピュータにライト終了を報告するものであり、中間コピー先ボリューム V 2 において、非同期リモートコピーで受領したリモートコピーデータを同期リモートコピーで他サイトのボリュームにリモートコピーすることは、あまり意味が無く現実的でないためである。

【0045】ライト処理 3 1 では、まず、ライト要求を受領し (ステップ S 4 0 1)、ライト要求元がホストコンピュータ 1、かつ、ライト対象ボリュームがリモートコピー元でないか確認する (ステップ S 4 0 2)。

【0046】ステップ S 4 0 2 の条件成立の場合、ホストコンピュータ 1 からリモートコピーに関連しないボリュームへのライト要求であるため (上記 1) に相当)、ライトデータをキャッシュメモリ 5 の通常領域に格納し (ステップ S 4 0 3)、ライト要求元にライト終了を報告する (ステップ S 4 0 4)。このライトデータは、ホストコンピュータ 1 からのライト要求とは非同期に、記憶制御装置 2 1 が記憶装置 2 2 に書き込む。

【0047】ステップ S 4 0 2 の条件不成立の場合、ライト要求元がホストコンピュータ 1 であるか確認する (ステップ S 4 0 5)。

【0048】ステップ S 4 0 5 の条件成立の場合 (上記 2) 3) 4) に相当)、ライトデータに対応させるライ

トシーケンス管理情報エントリ721を確保し、ライトシーケンス#を付与する。具体的には、ホストコンピュータ1からライトされたボリュームに対応するボリューム管理情報63のライトシーケンス管理情報#801を参照し、対応するライトシーケンス管理情報62のライトシーケンス管理情報エントリ721を確保し、ライトシーケンス#カウンタ71の値を設定し、ライトシーケンス#カウンタ71の値を加算し、ライトシーケンス管理情報エントリ721をライトデータと関連付ける。

【0049】ライトシーケンス#の付加は、リモートコピーの同期/非同期に関係なく実行する。これは、仮に当該ボリュームの直接のコピー先が同期リモートコピー先だけであったとしても、これらのコピー先のさらに先が非同期リモートコピーを実行することがあるためである。

【0050】ステップS405の条件不成立の場合（上記6）～10）に相当）、ライトデータに対応させるライトシーケンス管理情報エントリ721を確保し、ライトデータと共に受領したライトシーケンス#を含むライトシーケンス管理情報62を格納する（ステップS407）。

【0051】次に、ライトデータをキャッシュメモリ5に格納する。キャッシュメモリ5への格納方法は、ライト対象ボリュームにより異なり、A) 非同期リモートコピー先である（上記9）10）に相当）、B) 非同期リモートコピー先ではなく、非同期リモートコピー元である（上記3）4）7）8）に相当）、C) それら以外（上記2）5）6）に相当）の3通りに分けられる。これは、ライト対象ボリュームがA) である場合、受領したライトデータを一旦仮保存し、ライトシーケンス#に従って正式にライトする必要があるためである。また、B) である場合、非同期リモートコピー先に未転送のライトデータをライトされた場合、古いライトデータを転送するまで別に保存しておく必要があるためである。

【0052】まず、ライト対象ボリュームが非同期リモートコピー先かどうか判断する（ステップS408）。

【0053】ステップS408の条件成立（上記A）に相当）の場合、当該ライトデータはホストコンピュータ1からの更新順を示すライトシーケンス#に従ってライトする必要があるため、一旦、ライトデータをキャッシュメモリ5の仮領域に格納し（ステップS409）、ライトデータに対応するライトシーケンス管理情報エントリ721を正式化対象データとして正式化対象ライトシーケンス管理情報74に登録する（ステップS410）。図3においては、ライトシーケンス管理情報エントリ721\_\_e、721\_\_fが相当する。

【0054】さらに、ライト対象ボリュームが非同期リモートコピーのコピー元であるか判断し（ステップS411）、ステップS411の条件成立（上記3）4）7）8）10）に相当）の場合、ライトデータに対応す

るライトシーケンス管理情報エントリ721を転送対象データとして転送対象ライトシーケンス管理情報73に登録する（ステップS412）。図3においては、ライトシーケンス管理情報エントリ721\_\_b、721\_\_c、721\_\_d、721\_\_eが相当する。

【0055】ステップS408の条件成立の場合、さらに、ライト対象ボリュームが非同期リモートコピーのコピー元であるか判断する（ステップS413）。

【0056】ステップS413の条件成立（上記B）に相当）の場合、非同期リモートコピー先へ未転送のライトデータをキャッシュメモリ5の仮領域へ移設した上で、ライトデータをキャッシュメモリ5の通常領域に格納する（ステップS414）。図3においては、ライトシーケンス管理情報エントリ721\_\_b、721\_\_c、721\_\_dが相当する。さらに、ビットマップ更新処理32を実行する（ステップS415）。ビットマップ更新処理32については後述する。その後、ステップS411へ。

【0057】ステップS413の条件不成立（上記C）に相当）の場合、当該ライトデータは、非同期リモートコピーのコピー先でもコピー元でもないため、ライトデータをキャッシュメモリ5の通常領域に格納する（ステップS416）。図3においては、ライトシーケンス管理情報エントリ721\_\_aが相当する。さらに、ビットマップ更新処理32を実行する（ステップS417）。ビットマップ更新処理32については後述する。その後、ステップS418へ。

【0058】次に、ライト対象ボリュームが同期リモートコピーのコピー元であるか判断し（ステップS418）、ステップS418の条件成立の場合（上記2）4）6）8）に相当）、同期リモートコピーのコピー先にライトデータとライトシーケンス管理情報エントリ721の内容を転送する（ステップS419）。当該同期リモートコピーは、ライト要求元へのライト要求を受領してからライト終了を報告するまでの時間を短縮するために、複数のコピー先に対し並列に実行しても良い。

【0059】次に、当該ライトデータに対応するライトシーケンス管理情報721が、転送対象ライトシーケンス管理情報73、および、正式化対象ライトシーケンス管理情報74のいずれにも登録されていないか判断する（ステップS420）。ステップS420の条件成立（上記2）5）6）に相当）の場合、当該ライトデータに対応するライトシーケンス管理情報エントリ721は不要であるため、ライトシーケンス管理情報エントリ721を解放する（ステップS421）。図3においては、ライトシーケンス管理情報エントリ721\_\_gが相当する。

【0060】最後に、ライト要求元にライト終了を報告（ステップS404）し、処理を終了する。

【0061】図5に、記憶装置システム2の内、非同期

リモートコピーのコピー元ボリュームを持つ記憶装置システム2で実行する非同期転送処理33について説明する。当該処理は、ライトシーケンス管理情報62単位に実行しても良い。

【0062】まず、転送対象ライトシーケンス管理情報73に登録されたライトシーケンス管理情報エントリ721から転送対象を選択する(ステップS501)。

【0063】次に、非同期リモートコピー先に、選択したライトシーケンス管理情報エントリ721と対応するライトデータを転送する(ステップS502)。

【0064】次に、転送したライトシーケンス管理情報エントリ721の転送対象ライトシーケンス管理情報73への登録を解除する(ステップS503)。

【0065】次に、転送したライトシーケンス管理情報エントリ721が正式化対象ライトシーケンス管理情報74にも登録されているか判断する(ステップS504)。ステップS504の条件不成立の場合、非同期リモートコピー対象データとしての管理が不要となるため、転送したライトシーケンス管理情報エントリ721を解放し(ステップS505)、転送したライトデータがキャッシュメモリ5の仮領域を使用していた場合には、当該仮領域も解放する(ステップS506)。

【0066】図6に、記憶装置システム2の内、非同期リモートコピーのコピー先ボリュームを持つ記憶装置システム2で実行する非同期正式化処理41について説明する。当該処理は、ライトシーケンス管理情報62単位に実行しても良い。

【0067】まず、正式化対象ライトシーケンス管理情報74に登録されたライトシーケンス管理情報エントリ721を参照し、ライトシーケンス#が連続している範囲を正式化対象として選択する(ステップS601)。

【0068】次に、選択した正式化対象のライトデータは全て正式化済みか判断する(ステップS602)。

【0069】ステップS602の条件成立の場合、ステップS606へ。

【0070】ステップS602の条件不成立の場合、正式化対象として選択したライトデータの内、最も古いライトシーケンス#に対応するライトデータをキャッシュメモリ5の仮領域から通常領域に移設し、正式化する(ステップS603)。

【0071】次に、正式化したライトシーケンス管理情報エントリ721の正式化対象ライトシーケンス管理情報74への登録を解除する(ステップS604)。

次に、ビットマップ更新処理32を実行する(ステップS605)。ビットマップ更新処理32については後述する。その後、ステップS602へ戻る。

【0072】次に、正式化したライトシーケンス管理情報エントリ721が転送対象ライトシーケンス管理情報73にも登録されているか判断する(ステップS606)。ステップS606の条件不成立の場合、非同期リ

モートコピー対象データとしての管理が不要となるため、正式化したライトシーケンス管理情報エントリ721を解放し(ステップS607)、正式化したライトデータが使用していたキャッシュメモリ5の仮領域を解放する(ステップS608)。

【0073】図7に、記憶装置システム2の内、総コピー元ボリュームV1を持つ記憶装置システム2で実行するビットマップ切り替え起動処理34について説明する。当該処理は、ビットマップ有効フラグ810がONであるビットマップ809(2つ以上存在する場合には記録開始ライトシーケンス#811の古い方)の更新データ量カウンタ812の値が一定の閾値を超えた場合に、ボリューム単位に実行しても良い。

【0074】また、前回実行したビットマップ切り替えからの経過時間が一定の閾値を超えた場合に、ボリューム単位に実行しても良い。

【0075】まず、総コピー元ボリュームV1、および、当該ボリュームを総コピー元ボリュームV1とする全ての中間コピー先ボリュームV2と末端コピー先ボリュームV3に対して、前回のビットマップ切り替え終了前に新たにビットマップ切り替えを開始しないようにするため、ビットマップ切り替え中フラグ804がOFFであるか判断する(ステップS701)。

【0076】ステップS701の条件不成立の場合、まだビットマップ切り替え中のため、処理を終了する。

【0077】ステップS701の条件成立の場合、ビットマップ切り替え中フラグ804をONする(ステップS702)。

【0078】次に、ビットマップ切り替え契機とするライトシーケンス#を決定し、管理情報メモリのビットマップ切り替え契機ライトシーケンス#806に設定する(ステップS703)。ビットマップ切り替え契機とするライトシーケンス#の決定方法として、総コピー元ボリュームV1から、当該ボリュームを総コピー元ボリュームV1とする全ての中間コピー先ボリュームV2と末端コピー先ボリュームV3へ通知するまでに、中間コピー先ボリュームV2と末端コピー先ボリュームV3の最新ライトシーケンス#805が超えてしまわない程度に進んだライトシーケンス#を生成し、ビットマップ切り替え契機ライトシーケンス#806として良い。

【0079】また、ビットマップ切り替え契機ライトシーケンス#806が、切り替え先ビットマップ809の記録開始ライトシーケンス#となることを意図しているが、1つのライトシーケンス管理情報62で複数のボリュームのライトシーケンスを管理する場合、決定したビットマップ切り替え契機ライトシーケンス#806が必ずしも当該ボリュームのライトデータに対して割り当てられるとは限らない。しかし、決定したビットマップ切り替え契機ライトシーケンス#806を超えて当該ボリュームに最初に割り当てられるライトシーケンス#は、

総コピー元ボリュームV1と、当該ボリュームを総コピー元ボリュームV1とする全ての中間コピー先ボリュームV2と末端コピー先ボリュームV3において、一意に定まるため、ビットマップ切り替え契機ライトシーケンス#806を超えて当該ボリュームに最初に割り当てられるライトシーケンス#を、切り替え先ビットマップ809の記録開始ライトシーケンス#とすることができ

る。

【0080】次に、最新ライトシーケンス#805がビットマップ切り替え契機ライトシーケンス#806と同じとなった、あるいは、超えたことを知らせるライトシーケンス#到達フラグ808をOFFする（ステップS704）。

【0081】次に、ビットマップ更新処理32に最新ライトシーケンス#805とビットマップ切り替え契機ライトシーケンス#806の比較を要求するライトシーケンス#確認要求フラグ807をONする（ステップS705）。

【0082】次に、当該ボリュームの直接のコピー先全てに対し、ビットマップ切り替え契機ライトシーケンス#806を転送し、最新ライトシーケンス#がビットマップ切り替え契機ライトシーケンス#806を超えた場合のライトシーケンス#到達報告を要求する（ステップS706）。

【0083】次に、当該ボリュームのライトシーケンス#到達フラグ808がON、かつ、当該ボリュームの直接のコピー先全てからライトシーケンス#到達報告が有るか判断する（ステップS707）。ステップS707の条件不成立の場合、一定時間待ち（ステップS708）、再びステップS707の判断を実行する。なお、当該ボリュームの直接のコピー先からの報告は、逐次、コピー先情報802に記録するものとする。

【0084】ステップS707の条件成立の場合、自サイトのビットマップ809の内、記録開始ライトシーケンス#811が古い方の有効フラグ810をOFFし、当該ビットマップ809をクリアする（ステップS709）。これは、この時点で、当該ボリュームと、当該ボリュームを総コピー元ボリュームV1とする全ての中間コピー先ボリュームV2と末端コピー先ボリュームV3の最新ライトシーケンス#805がビットマップ切り替え契機ライトシーケンス#806を超え、新たなビットマップ809への記録を開始済みであるためである。

【0085】次に、当該ボリュームの直接のコピー先全てに対し、ビットマップ切り替えを要求する（ステップS710）。

【0086】次に、当該ボリュームの直接のコピー先全てからビットマップ切り替え完了報告があるか判断する（ステップS711）。ステップS711の条件不成立の場合、一定時間待ち（ステップS712）、再びステップS711の判断を実行する。

【0087】ステップS711の条件成立の場合、ビットマップ切り替え終了のため、ビットマップ切り替え中フラグ804をOFFし（ステップS713）、処理を終了する。

【0088】図8に、記憶装置システム2の内、中間コピー先ボリュームV2と末端コピー先ボリュームV3を持つ記憶装置システム2で実行するビットマップ切り替え処理35について説明する。

【0089】まず、コピー元サイトからビットマップ切り替え契機ライトシーケンス#806を受領し、ビットマップ切り替え対象となる中間コピー先ボリュームV2と末端コピー先ボリュームV3のビットマップ切り替え契機ライトシーケンス#806に設定する（ステップS801）。

【0090】次に、ライトシーケンス#到達フラグ808をOFFし（ステップS802）、ライトシーケンス#確認要求フラグ807をONする（ステップS803）。

【0091】次に、当該ボリュームに直接のコピー先が存在するか判断する（ステップS804）。

【0092】ステップS804の条件不成立（末端コピー先ボリュームV3に相当）の場合、ステップS808へ。

【0093】ステップS804の条件成立（中間コピー先ボリュームV2に相当）の場合、当該ボリュームの直接のコピー先全てに対し、ビットマップ切り替え契機ライトシーケンス#806を転送し、最新ライトシーケンス#805がビットマップ切り替え契機ライトシーケンス#806を超えた場合のライトシーケンス#到達報告を要求する（ステップS805）。

【0094】次に、当該ボリュームの直接のコピー先全てからライトシーケンス#到達報告が有るか判断する（ステップS806）。ステップS806の条件不成立の場合、一定時間待ち（ステップS807）、再びステップS806の判断を実行する。

【0095】ステップS806の条件成立の場合、当該ボリュームのライトシーケンス#到達フラグがONであるか判断する（ステップS808）。ステップS808の条件不成立の場合、一定時間待ち（ステップS809）、再びステップS808の判断を実行する。

【0096】ステップS808の条件成立の場合、当該ボリュームの直接のコピー元にライトシーケンス#到達を報告する（ステップS810）。

【0097】次に、当該ボリュームの直接のコピー元からビットマップ切り替え要求を受領したか判断する（ステップS811）。ステップS811の条件不成立の場合、一定時間待ち（ステップS812）、再びステップS811の判断を実行する。

【0098】ステップS811の条件成立の場合、当該ボリュームのビットマップ809の内、記録開始ライト

シーケンス#811が古い方のビットマップ有効フラグ810をOFFし、当該ビットマップ809をクリアする(ステップS813)。

【0099】次に、当該ボリュームに直接のコピー先が存在するか判断する(ステップS814)。

【0100】ステップS814の条件不成立(末端コピー先ボリュームV3に相当)の場合、ステップS818へ。

【0101】ステップS814の条件成立(中間コピー先ボリュームV2に相当)の場合、当該ボリュームの直接のコピー先全てに対し、ビットマップ切り替えを要求する(ステップS815)。

【0102】次に、当該ボリュームの直接のコピー先全てからビットマップ切り替え完了報告の有るか判断する(ステップS816)。ステップS816の条件不成立の場合、一定時間待ち(ステップS817)、再びステップS816の判断を実行する。

【0103】ステップS816の条件成立の場合、当該ボリュームの直接のコピー元にビットマップ切り替え完了を報告し(ステップS818)、処理を終了する。

【0104】図9に、ライト処理31、非同期正式化処理41で実行するビットマップ更新処理32について説明する。

【0105】まず、ビットマップ有効フラグ810がONであるビットマップ809のライトデータに対応するビットをONし、更新したビットマップ809に対応した更新データ量カウンタ812をインクリメントする(ステップS901)。

【0106】次に、最新ライトシーケンス#805を当該ライトデータに対応したライトシーケンス#に更新する(ステップS902)。

【0107】次に、ライトシーケンス#確認要求フラグ807がONであるか判断する(ステップS903)。

【0108】ステップS903の条件不成立の場合、処理を終了する。

【0109】ステップS903の条件成立の場合、最新ライトシーケンス#805がビットマップ切り替え契機ライトシーケンス#806と同じ、あるいは、超えるか判断する(ステップS904)。

【0110】ステップS904の条件不成立の場合、処理を終了する。

【0111】ステップS904の条件成立の場合、ライトシーケンス#到達フラグ808をONする(ステップS905)。

【0112】次に、ビットマップ有効フラグ810がOFFであるビットマップ809のビットマップ有効フラグ810をONし、記録開始ライトシーケンス#811に最新ライトシーケンス#805を設定する(ステップS906)。

【0113】次に、新たにビットマップ有効フラグ81

0をONしたビットマップ809におけるライトデータに対応するビットをONし、更新したビットマップ809に対応した更新データ量カウンタ812をインクリメントする(ステップS907)。

【0114】最後に、ライトシーケンス#確認要求フラグ807をOFFし(ステップS908)、処理を終了する。

【0115】図10に、ビットマップ切り替えにおける、各サイトの処理とサイト間の通信の概要を示す。

【0116】総コピー元ボリュームV1を持つサイトで実行しているのがビットマップ切り替え起動処理34、中間コピー先ボリュームV2、または、末端コピー先ボリュームV3を持つサイトで実行しているのがビットマップ切り替え処理35である。ただし、一部ビットマップ更新処理32を含む。各実行内容には、該当するステップ番号を付加した。

【0117】図10は、サイト全体のビットマップ1からビットマップ2へのビットマップ切り替えの概要を示しており、ビットマップ切り替えは、この後、ビットマップ2からビットマップ1へ、さらに、ビットマップ1からビットマップ2へとといったように、交互に実行する。

【0118】ビットマップ切り替えを、サイト全体として、単一のビットマップを更新する状態から、一時的に2重のビットマップを更新する状態を経て、再び単一のビットマップを更新する状態へ移行することにより、ビットマップ切り替え中を含む任意の瞬間、任意のサイト同士で必ず記録開始ライトシーケンス#811が同じビットマップ809を共有することができ、ライトシーケンス#の進んだサイト側にボリューム内容を一致させることができる。

【0119】また、ビットマップ切り替えにおいて、一部のサイトの障害を検出した場合、ビットマップ切り替え契機ライトシーケンス#806の各サイトの記憶装置システム2への転送が失敗することがある。この場合、総コピー元ボリュームV1を持つ記憶装置システム2において、ステップS709のビットマップ809クリアの内容を実行せず、ステップS710において、ビットマップ809をクリアせずビットマップ切り替え完了を報告するよう他サイトの記憶装置システム2に要求することで、障害を発生したサイトも含めた差分の管理を続行することができる。

【0120】さらに、各記憶装置システム2においてビットマップ809を3重以上持ち、一部サイトの障害を検出した場合、上記のように、障害を検出した時点で使用していたビットマップ809をクリアしないようにし、残りのビットマップを使用してビットマップ切り替えを続行することで、障害を発生したサイトも含めた差分の管理を続行しながら、正常なサイト間の差分の増大を抑止することができる。クリアしないようにしたビッ

ットマップ809を、すべてのサイトが正常となるまで維持することで、障害サイトがさらに増えても、障害が発生したサイトも含めた差分の管理を続行することができる。

【0121】また、ビットマップ切り替えに伴う記憶装置システム2間の通信内容を、通常のリモートコピーデータの転送、および、応答に添付して実行することで、サイト間の通信負荷を削減できる。

【0122】図11に、サイト被災後、サイト間でボリューム内容を一致させるための手順について説明する。

【0123】まず、サイト被災により直接のコピー元を失った自ボリュームの新たな直接のコピー元として、自ボリュームと同じ総コピー元ボリュームV1を持つボリュームの内、自ボリュームより最新ライトシーケンス#805が進んでいるボリュームを選択する（ステップS1101）。

【0124】次に、自サイトと新たな直接のコピー元サイトの記憶装置システム2間に転送パスを設定する（ステップS1102）。

【0125】次に、自サイトと新たな直接のコピー元サイトの記憶装置システム2間でリモートコピーを開始する（ステップS1103）。これ以降、新たな直接のコピー元サイトのコピー元ボリュームに対してなされた更新は、自サイトのコピー先ボリュームに逐次反映される。

【0126】最後に、自サイトと新たな直接のコピー元サイトの記憶装置システム2間で差分コピー処理36を起動する（ステップS1104）。差分コピー処理36については後述する。

【0127】差分コピー処理36終了後も、そのままリモートコピーを継続することで、通常と同じリモートコピー動作となる。

【0128】上記の新たな直接のコピー元サイトとして、あらかじめ、自ボリュームについて自サイトより総コピー元ボリュームV1に近い上流側のサイトを登録しておいても良い。

【0129】図12に、サイト被災後、サイト間でボリューム内容を一致させるための手順の1つである差分コピー処理36について説明する。

【0130】まず、新たな直接のコピー元サイトにおいて、自サイトと新たな直接コピー元サイトの両サイトでビットマップ有効フラグ810がONであるビットマップ809で、かつ、この条件を満たすビットマップが複数存在する場合には、記録開始ライトシーケンス#811が新しい方のビットマップ809を選択し、ビットマップ809の内容を、差分コピー処理36でコピーするデータを決定するための差分ビットマップ813にコピーする（ステップS1201）。差分ビットマップ813にコピーする理由は、新たな直接のコピー元サイトでは、サイト被災後もリモートコピーが継続している場合

があり、ビットマップ有効フラグ810がONであるビットマップ809が更新され続けることがあるため、差分コピー量の増大を防ぐため、現状の両サイトのボリューム間の差分を保存しておく必要があるためである。

【0131】次に、差分ビットマップ813でONとなっているビットに対応する部分を、新たな直接のコピー元サイトのコピー元ボリュームから自サイトのコピー先ボリュームにコピーする（ステップS1202）。

【0132】最後に、差分ビットマップ813をクリアし（ステップS1203）、処理を終了する。

【0133】

【発明の効果】本発明によれば、主に3サイト以上のNサイトに配置した記憶制御装置間のリモートコピーにおいて、1サイト被災後、速やかに残サイト間のリモートコピーデータの一致を実現するための差分を管理することができる。また、リモートコピーを構成するサイト数Nに依存せず、差分を管理する情報量を一定に保つことができる。

【図面の簡単な説明】

【図1】本発明の実施の形態における、Nサイト間のリモートコピーの構成例を示す。

【図2】本発明の実施の形態における、記憶装置システムの構成例を示す。

【図3】本発明の実施の形態における、ライトシーケンス管理情報とキャッシュメモリの関係の一例を示す。

【図4】本発明の実施の形態における、記憶装置システムで実行するライト処理の一例を示す。

【図5】本発明の実施の形態における、非同期リモートコピー元の記憶装置システムで実行する非同期転送処理の一例を示す。

【図6】本発明の実施の形態における、非同期リモートコピー先の記憶装置システムで実行する非同期正式化処理の一例を示す。

【図7】本発明の実施の形態における、総コピー元ボリュームV1を持つ記憶装置システムで実行するビットマップ切り替え起動処理の一例を示す。

【図8】本発明の実施の形態における、中間／末端コピー元ボリュームを持つ記憶装置システムで実行するビットマップ切り替え処理の一例を示す。

【図9】本発明の実施の形態における、ライト処理、および、非同期正式化処理で実行するビットマップ更新処理の一例を示す。

【図10】本発明の実施の形態における、ビットマップ切り替えの概要を示す。

【図11】本発明の実施の形態における、サイト被災後、サイト間でボリューム内容を一致させるための手順の一例を示す。

【図12】本発明の実施の形態における、サイト被災後、サイト間でボリューム内容を一致させるための手順の1つである差分コピー処理の一例を示す。

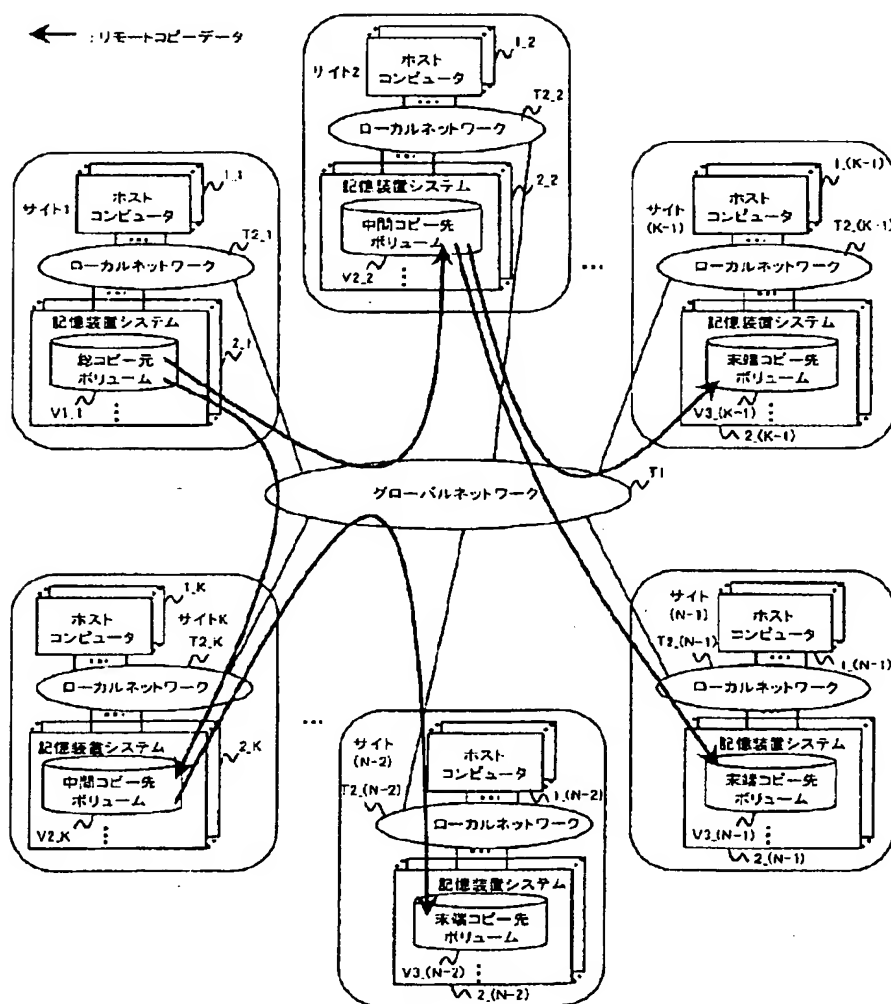
【符号の説明】

1…ホストコンピュータ、2…記憶装置システム、21…記憶制御装置、22…記憶装置、23…保守端末、3…チャネルインタフェース、31…ライト処理、32…ビットマップ更新処理、33…非同期転送処理、34…ビットマップ切り替え起動処理、35…ビットマップ切り替え処理、36…差分コピー処理、4…ディスクインタフェース、41…非同期正式化処理、5…キャッシュメモリ、6…管理情報メモリ、62…ライトシーケンス管理情報、63…ボリューム管理情報、71…ライトシーケンス#カウンタ、721…ライトシーケンス管理情報エントリ、73…転送対象ライトシーケンス管理情報、

74…正式化対象ライトシーケンス管理情報、804…ビットマップ切り替え中フラグ、805…最新ライトシーケンス#、806…ビットマップ切り替え契機ライトシーケンス#、807…ライトシーケンス#確認要求フラグ、808…ライトシーケンス#到達フラグ、809…ビットマップ、810…ビットマップ有効フラグ、811…ビットマップ記録開始ライトシーケンス#、812…ビットマップ更新量データカウンタ、813…差分ビットマップ、T1…グローバルネットワーク、T2…ローカルネットワーク、V1…総コピー元ボリューム、V2…中間コピー先ボリューム、V3…末端コピー先ボリューム。

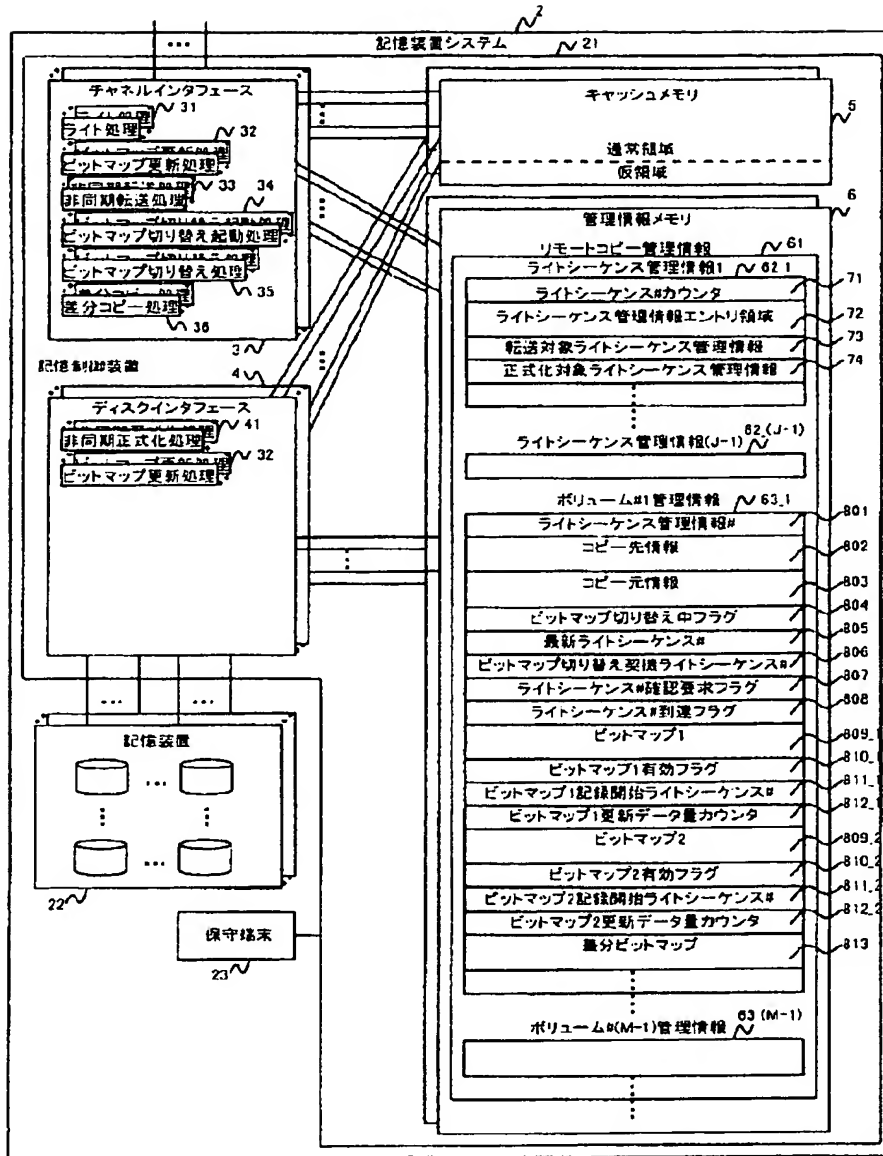
【図1】

図 1



【図2】

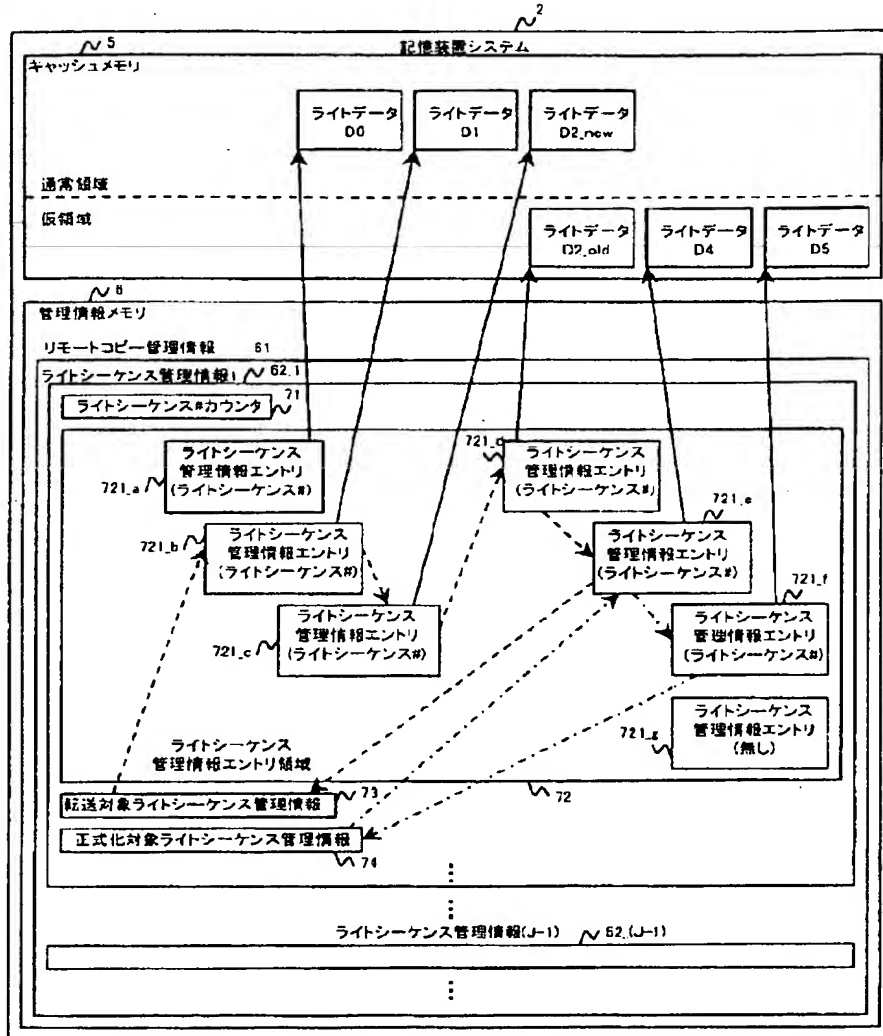
図 2





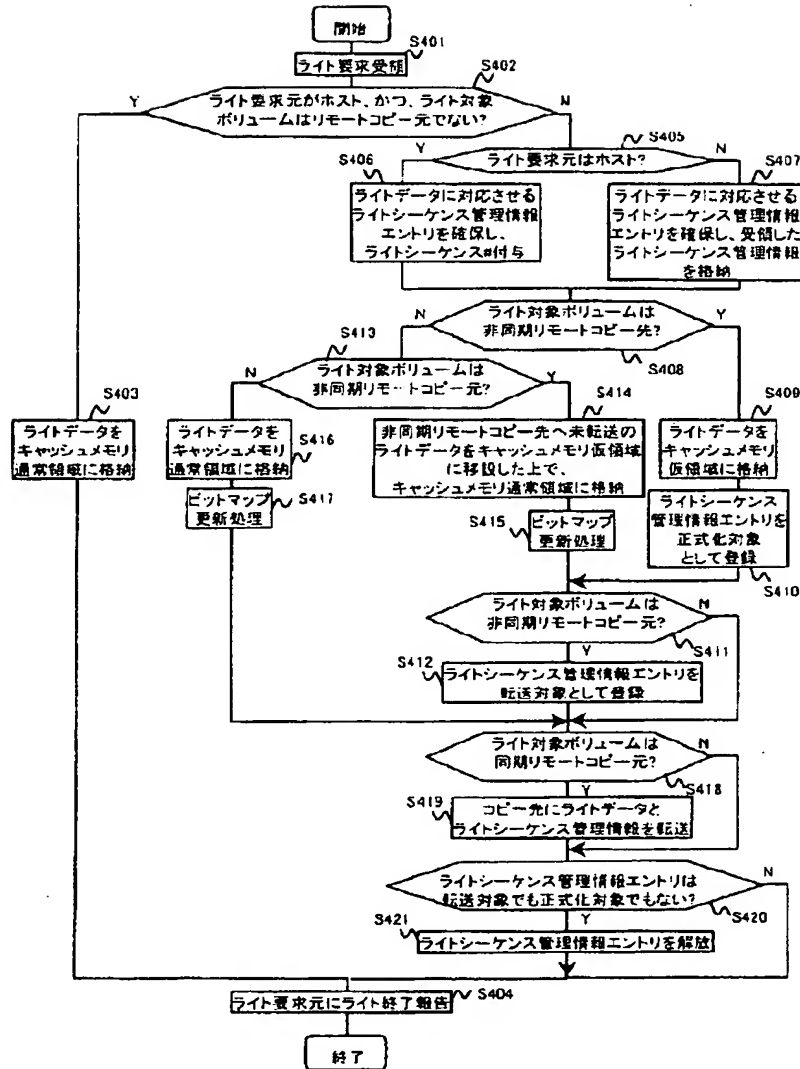
【図 3】

図 3



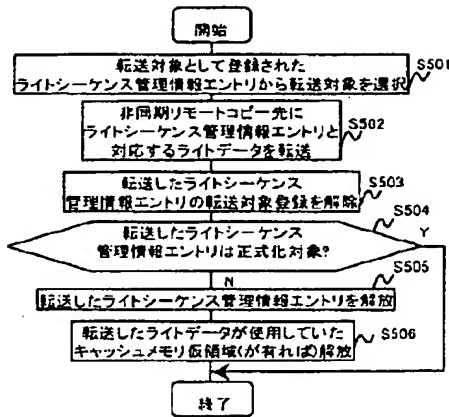
【図 4】

図 4



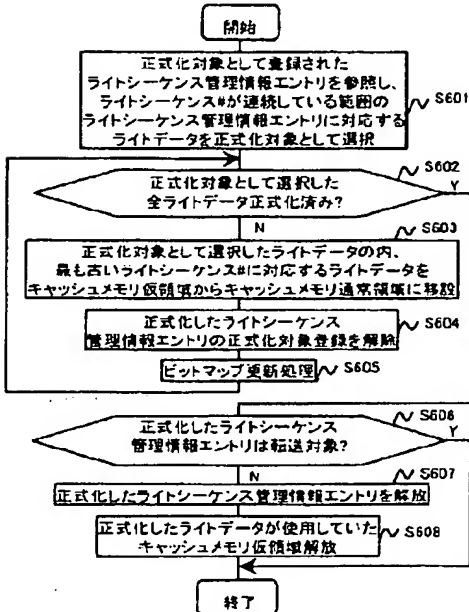
【図5】

図 5



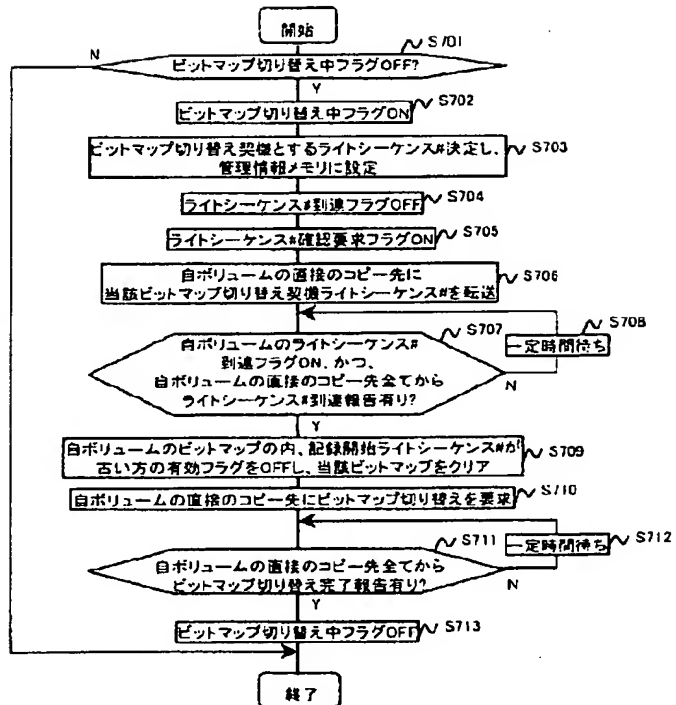
【図6】

図 6



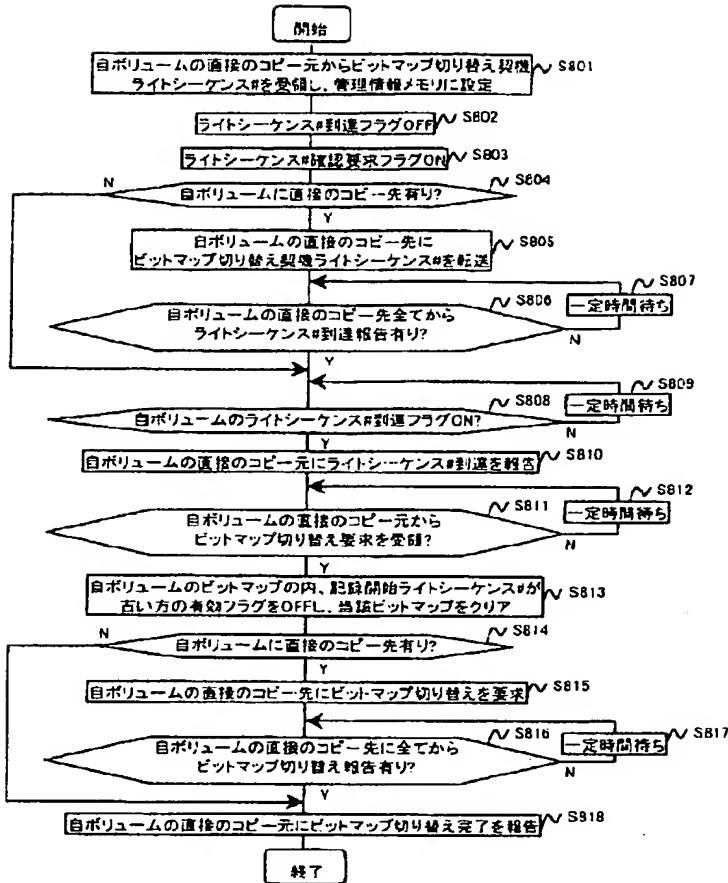
【図7】

図 7



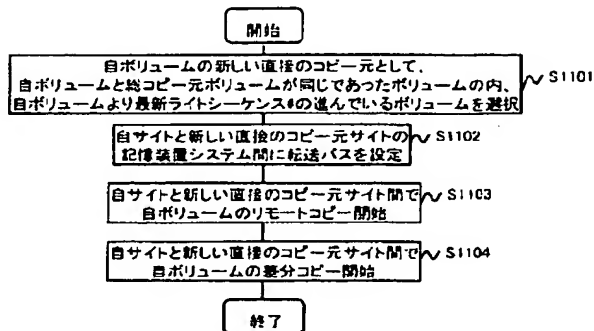
【図 8】

図 8



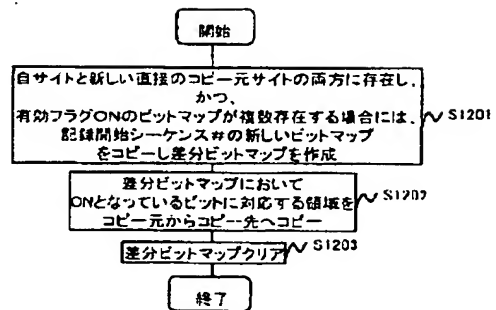
【図 11】

図 11



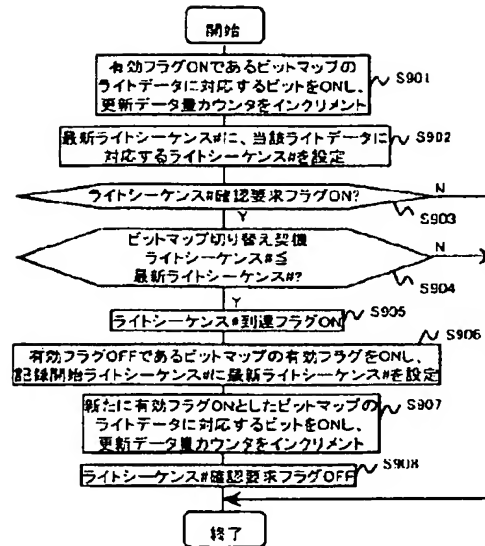
【図 12】

図 12



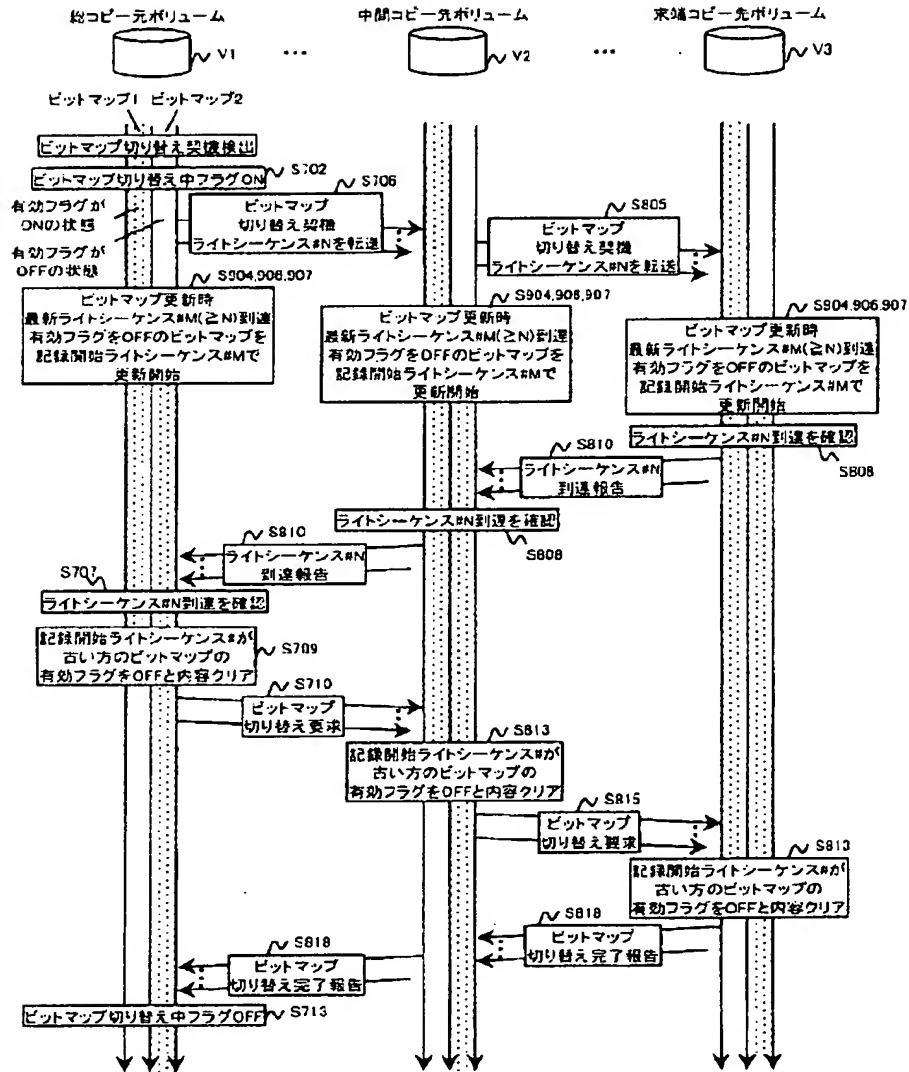
【図9】

図 9



【図10】

図 10



フロントページの続き

(72) 発明者 檜垣 誠一

神奈川県小田原市中里322番地2号 株式  
会社日立製作所RAIDシステム事業部内

Fターム(参考) 5B065 BA01 CC08 CE06 EA24 EA35

EK05

5B082 DC09 HA03

**This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record.**

## **BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☒ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☒ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** \_\_\_\_\_

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**